TURUN YLIOPISTON JULKAISUJA
ANNALES UNIVERSITATIS TURKUENSIS

SARJA - SER. D OSA - TOM. 918

MEDICA - ODONTOLOGICA

GENETIC REGULATORY NETWORKS IN AVIAN B CELLS

by

Pekka Kohonen

TURUN YLIOPISTO
UNIVERSITY OF TURKU
Turku 2010

From the Turku Graduate School of Biomedical Sciences
Department of Medical Microbiology and Immunology, University of Turku

Supervised by

Professor Olli Lassila, MD PhD
Department of Medical Microbiology and Immunology
University of Turku
FI-20520, Turku, Finland

Reviewed by

Professor Seppo Meri,
Department of Bacteriology and Immunology
P.O. Box 21 (Haartmaninkatu 3)
University of Helsinki
FI-00014, Helsinki, Finland

and

Professor Mark Johnson
Department of Biosciences,
Åbo Akademi University
FI-20520, Åbo, Finland

Opponent

Professor Mauno Vihinen
Institute of Medical Technology, Bioinformatics
University of Tampere
FI-33014, Tampere, Finland

ISBN 978-951-29-4381-4 (PDF)
ISBN 978-951-29-4383-8 (PRINT)
ISSN 0355-9483
Painosalama Oy – Turku, Finland 2010

ABSTRACT

Pekka Kohonen

Genetic regulatory networks in avian B cells

From Turku Graduate School of Biomedical Sciences (TUBS), Department of Medical Microbiology and Immunology, University of Turku, Turku, Finland.

Biology is turning into an information science. The science of systems biology seeks to understand the genetic networks that govern organism development and functions. In this study the chicken was used as a model organism in the study of B cell regulatory factors. These studies open new avenues for plasma cell research by connecting the down regulation of the B cell gene expression program directly to the initiation of plasma cell differentiation.

The unique advantages of the DT40 avian B cell model system, specifically its high homologous recombination rate, were utilized to study gene regulation in Pax5 knock out cell lines and to gain new insights into the B cell to plasma cell transitions that underlie the secretion of antibodies as part of the adaptive immune response. The Pax5 transcription factor is central to the commitment, development and maintenance of the B cell phenotype. Mice lacking the Pax5 gene have an arrest in development at the pro-B lymphocyte stage while DT40 cells have been derived from cells at a more mature stage of development. The DT40 Pax5-/- cells exhibited gene expression similarities with primary chicken plasma cells. The expression of the plasma cell transcription factors Blimp-1 and XBP-1 were significantly upregulated while the expression of the germinal centre factor BCL6 was diminished in Pax5-/- cells, and this alteration was normalized by Pax5 re-introduction. The Pax5-deficient cells further manifested substantially elevated secretion of IgM into the supernatant, another characteristic of plasma cells. These results for the first time indicated that the downregulation of the Pax5 gene in B cells promotes plasma cell differentiation.

Cross-species meta-analysis of chicken and mouse Pax5 gene knockout studies uncovers genes and pathways whose regulatory relationship to Pax5 has remained unchanged for over 300 million years.

Restriction of the hematopoietic stem cell fate to produce T, B and NK cell lineages is dependent on the *Ikaros* and its molecular partners, the closely related *Helios* and *Aiolos*. *Ikaros* family members are zinc finger proteins which act as transcriptional repressors while helping to activate lymphoid genes. *Helios* in mice is expressed from the hematopoietic stem cell level onwards, although later in development its expression seems to predominate in the T cell lineage. This study establishes the emergence and sequence of the chicken *Ikaros* family members. *Helios* expression in the bursa of Fabricius, germinal centres and B cell lines suggested a role for Helios in the avian B-cell lineage, too.

Phylogenetic studies of the Ikaros family connect the expansion of the Ikaros family, and thus possibly the emergence of the adaptive immune system, with the second round of genome duplications originally proposed by Ohno. Paralogs that have arisen as a result of genome-wide duplications are sometimes termed ohnologs – Ikaros family proteins appear to fit that definition.

This study highlighted the opportunities afforded by the genome sequencing efforts and somatic cell reverse genetics approaches using the DT40 cell line. The DT40 cell line and the avian model system promise to remain a fruitful model for mechanistic insight in the post-genomic era as well.

Keywords: Pax5, Helios, B cell, plasma cell, adaptive immunity, evolution, microarrays, bioinformatics, network biology, meta-analysis

TIIVISTELMÄ

Pekka Kohonen

Geenisäätelyverkostot kanan B-soluilla

Turun biolääketieteellinen tutkijakoulu (TUBS), Lääketieteellinen Mikrobiologia ja Immunologia, Turun Yliopisto, Turku, Suomi.

Biologia on muuttumassa informaatiotieteeksi. Systeemibiologia pyrkii ymmärtämään geenien säätelyverkostoja, jotka määrittävät eliöiden kehitystä ja toimintoja. Tässä tutkimuksessa kanaa käytettiin mallina B-solujen säätelijöitä tutkittaessa. Tutkimus avasi uusia näkymiä vasta-aineita tuottavien solujen ymmärtämiseksi osoittamalla, että Pax5-geeninsäätelytekijän poistaminen B-soluilla johtaa vasta-aineita tuottavien solujen kehittymiseen.

Kanan DT40 B-solulinjan erityisominaisuutta, korkeaa DNA:n homologista integraatio-frekvenssiä, hyödynnettiin Pax5:n poistogeenisten solulinjojen tuottamiseksi, jotta voitiin tutkia B-solujen muuttumista plasmasoluiksi. Plasmasolut tuottavat vasta-aineita osana opittua eli adaptiivista immuunipuolustusjärjestelmää. Pax5-geeninsäätelijä on keskeisessä asemassa B-solujen erilaistumisen, kehityksen sekä ylläpidon kannalta. Hiiret, joilta puuttuu Pax5, eivät pääse pro-B-vaihetta pidemmälle B-solujensa kehityksessä. DT40-solut ovat jo ohittaneet tämän vaiheen, joten niiden avulla on mahdollista tutkia myöhempiä vaiheita. DT40 Pax5-poistogeeniset solut ilmensivät plasmasoluille tyypillisiä geenejä: Blimp1- ja XBP1-geeninsäätelijöiden ilmenemistasot nousivat voimakkaasti samalla, kun itukeskuskuksille ominainen BCL6-geeni lakkasi toimimasta. Pax5:n palauttaminen pyörsi tämän vaikutuksen. Pax5-poistogeeniset solut erittivät myös suuria määriä immuuniglobuliinia. Näin ollen voitiin todeta ensimmäisen kerran, että Pax5:n poisto aikaansai B-solujen muuttumisen plasmasoluiksi.

Lajirajoja ylittävä kanalla ja ihmisellä tehtyjen Pax5-poistogeenisten kokeiden tulosten meta-analyysi geenien ilmentymisen tasolla paljastaa keskeisiä säätelyverkostoja ja -reittejä, jotka ovat säilyneet muuttumattomina yli 300 miljoonan vuoden ajan.

Veren kantasolujen kehitysvaihtoehtojen kaventuminen, kunnes niistä tulee erilaistuneita T-, B- tai NK-soluja, riippuu Ikaros-perheen geeninsäätelijöistä: Ikaroksesta, Helioksesta ja Aioloksesta. Nämä sinkkisormiproteiinit auttavat aktivoimaan ja myös sammuttavat geenejä. Helioksen on todettu ilmenevän hiirillä lähinnä kantasoluissa sekä T-solulinjassa. Tässä tutkimuksessa todettiin kanojen ilmentävän Heliosta lisäksi erityyppisissä B-solupopulaatioissa, kuten bursan follikkeleissa ja B-solulinjoissa.

Ikaros-perheen evoluutiota tutkittiin fylogenetiikan menetelmillä. Synteenisiä genomisegmenttejä vertailemalla ja sekvenssi-analyysin avulla havaittiin, että Ikaros-perhe on todennäköisesti kahdentunut toisen genominlaajuisen kahdentumisvaiheen aikana (2R) – eli näitä geenejä voi kutsua ohnologeiksi. Adaptiivinen immuunijärjestelmä kehittyi samaan aikaan, ja näin ollen Ikaros-perheen synty voi liittyä oppivan immuunijärjestelmän syntyyn.

Tutkimuksessa hyödynnettiin genomisia menetelmiä ja DT40-poistogeenistä solulinja-mallia. Tämä mallisysteemi sopii erityisen hyvin mekanistisiin tutkimuksiin myös post-genomisella aikakaudella.

Avainsanoja: Pax5, Helios, B-solu, plasmasolu, oppiva immuunijärjestelmä, evoluutio, geenisirut, bioinformatiikka, verkostobiologia, meta-analyysi.

CONTENTS

ABBREVIATIONS

2R	second round of genome-wide duplication
AID	activation-induced cytidine deaminase
ATF6	activating transcription factor 6
BCL6	B cell lymphoma 6
BCP	B cell progenitor
BCR	B-cell receptor
bHLH	basic helix-loop-helix
Blimp-1	B lymphocyte-induced maturation protein 1 (*Prdm1* gene product)
BLNK	B cell linker protein
BSAP	B-cell specific activator protein
C/EBP	CCAAT/enhancer binding protein
CLP	common lymphoid progenitor
CSR	class switch recombination
cr-1	The chicken cr-1 repeat element
DAG	Directed acyclic graph
EBF	early B cell factor
ER	endoplasmic reticulum
ETP	early thymocute progenitor
FC	fold change
FDR	false discovery rate: the rate that false discoveries occur
FLT3	fms-related tyrosine kinase 3
FWER	familywise error rate
GC	germinal center
GCT	Gene Class Testing
GEO	Gene Expression Omnibus
GO	gene ontology
GRN	Gene Regulatory Network
GSA	Gene Set Analysis
GSEA	Gene Set Enrichment Analysis
HDAC	histone deacetylase complex
HLH	helix-loop-helix
HSC	hematopoietic stem cell
IGA	Individual gene analysis
IgH	immunoglobulin heavy chain
IgL	immunoglobulin light chain
IL-7	interleukin 7
IL-7Rα	interleukin receptor 7α
IRE1	inositol-requiring enzyme 1
IRF4	interferon regulatory factor 4
IRF8	interferon regulatory factor 8
ITAM	immunoreceptor tyrosine-based activation motif
kb	kilobase
LFC	log fold change (log in base 2)
Lin	lineage marker

LLPC	long-lived plasma cells
LMPP	lymphoid-myeloid multipotent progenitor
loess	locally weighted scatterplot smoothing
LPS	lipopolysaccharide
mIgμ	membrane form of immunoglobulin μ heavy chain
NK cell	natural killer cell
NURD complex	nucleosome remodelling and deacetylation complex
OTU	operational taxonomic unit
PAM	Point accepted mutation
Pax5	paired box protein 5
PCR	polymerace chain reaction
pFDR	positive FDR: the rate that discoveries are false
pre-BCR	pre-B-cell receptor
q-RT-PCR	quantitative RT-PCR
RAG	recombination-activating gene
RT-PCR	reverse transcription polymerase chain reaction
SAM	significance analysis of microarrays
SHM	somatic hypermutation
sIgM	surface immunoglobulin M
siRNA	small interfering RNA
SHP-1	Src-domain-2-containing protein tyrosine phosphatise
STRC	short term reconstituting stem cells
T_{FH}	follicular helper T cells
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
UPR	unfolded protein response
UV	ultra-violet light
VSN	variance stabilizing normalization
VST	variance stabilizing transfromation
WGD	Whole Genome Duplication
XBP1	X-box binding protein 1

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original publications, which are referred to in the text by the Roman numerals (I-IV).

I Nera KP, Kohonen P, Narvi E, Peippo A, Mustonen L, Terho P, Koskela K, Buerstedde JM, Lassila O. Loss of Pax5 promotes plasma cell differentiation. Immunity. 2006; 24(3):283-93.

II Kohonen P, Nera KP, Lassila O. Avian Helios and evolution of the Ikaros family. Scand J Immunol. 2004; 60(1-2):100-7.

III Kohonen P, Nera KP, Lassila O. Avian model for B-cell immunology--new genomes and phylotranscriptomics. Scand J Immunol. 2007; 66(2-3):113-21.

IV Kohonen P, Alinikula J, Nera KP, Lassila O. Cross-species meta-analysis of regulatory networks of the Pax5 transcription factor. 2010. Manuscript.

1. INTRODUCTION

A metazoan genome can be thought of as one of the most densely packaged forms of information in existence - and the development of an organism as a way to unravel that information. However, the genome is more like a cooking recipe than a blueprint of how to make, say, a human. A genome does not exist outside the context of the cell, organ, body or even the environment it resides in – which makes it impossible to understand how the various structures and functions it encodes come about, by looking at the genomic sequence data alone. Experiments, done either by man or by evolution, can pinpoint which parts of the genome appear most important for a given phenomenon. An approach to biological investigation, which calls itself systems biology, seeks to combine different kinds of data and various evolutionary timescales in order to understand what governs the organism's development and functions in both health and disease. An ultimate goal is to be able to model these processes in detail.

The core subject matter of this thesis deals with the details of how a cell, specifically the avian B cell, changes its programming: which factors are important and which mechanism are employed. The hematopoietic or blood forming system has served as a model for both stem cell biology as well as systems biology, especially as regards genetic regulatory networks in cellular differentiation. This is largely due to the easy accessibility of blood cells and lymphoid cell in their own compartment to experimentation, making it possible to characterize in detail the various cell populations which form the intermediate stages in the process that leads to the formation of the mature immune system.

2. REVIEW OF THE LITERATURE

2.1. B cell development from stem cells to mature B cells

Hematopoietic stem cell (HSC) differentiation is a paradigm for the development of tissue specific stem cells (Bryder et al. 2006; Orkin 2000). Differentiation of the hematopoietic stem cell populations into effector cells, including B and T cells, is a stepwise process involving a progressive restriction of the cell fate and proliferation potential (Miyamoto et al. 2002; Hu et al. 1997; Ivanova et al. 2002; Laslo et al. 2006). The development of hematopoietic stem cells into specialized cells is an ongoing process throughout the life span of an organism, although there are significant differences in the sites and to some extent in the details of this process during embryonic and adult life. The particulars of hematopoietic stem cell development are also well conserved, not just across mammals but also between the avian species and mammals (Lassila et al. 1978; Eilken et al. 2009; Boisset et al. 2010; Durand et al. 2005).

2.1.1. HSCs, LMPPs and multilineage priming

Hematopoietic stem cells go through many steps or stages before committing to the B-cell lineage (Hardy et al. 2007; Ramírez et al. 2010; Durand et al. 2005). According to current thinking, HSC differentiation splits at an early phase into Lymphoid-myeloid and Erythroid branches of differentiation (Yoshida et al. 2010). In the Lymphoid-myeloid branch, steps just prior to the B-cell stage include lymphoid multipotent progenitor cells (LMPP) (Adolfsson et al. 2005).

In addition to stem-cell specific genes the HSC cells express, at a low level, regulatory genes important for various specific lineages such as the myeloid, lymphoid fates and erythroid lineages (Ivanova et al, 2002, Miyamato et al, 2002). This has been termed multilineage priming in stem cells (Miyamoto 2002; Hu et al. 1997; Ivanova et al. 2002; Laslo et al. 2006, Yoshida et al. 2010). A specific chromatin state, called the bivalent state, was characterized that coincides with many transcription factor genes that are expressed at lower level (Bernstein et al., 2006). In molecular terms, this state involves H3 lysine-27 methylated regions interspersed with H3 lysine-4 methylation. Random or stochastic activation of genetic programs is involved in lineage specification during hematopoietic cell differentiation (Yoshida et al. 2010).

A signature of stochastic priming might be allelic exclusion or the expression of a key regulatory gene only from one allele in any given single cell within a defined cell population (Nutt & Busslinger 1999). Recent high-throughput sequencing has revealed that monoallelic expression of genes is surprisingly common in B cells and otherwise (Gimelbrant et al. 2007). This event could then be enhanced via positive feedback to increase the rate of transcription initiation, and possibly to activate both of the alleles.

Subsequently, the allele-specific pattern of expression can be passed onto successive generations via epigenetic mechanisms.

2.1.2. Priming of the lymphoid program

The transcription factors PU.1, E2A and Ikaros lay the foundation for the lymphoid differentiation program and eventually the commitment to B cell development (Ramírez et al. 2010) (Figure 1). The pre-B-cell receptor and B-cell receptor expression, their generation via recombination and their diversification to form fully a mature B-cell receptor repertoire play a crucial role in the development of B cells (Figure 2). Recently, the co-operative actions of E2A and Foxo1 have been linked to early specification and induction of EBF1 and hence B-cell lineage specific transcription (Rothenberg 2010; Lin et al. 2010; Bryder & Sigvardsson 2010).

Ikaros is an "early acting" epigenetic regulator that establishes the partly open or primed bivalent chromatin states characterized by H3K4 mono-, di- or trimethylation (Ramírez et al. 2010; Ng et al. 2009). Ikaros, in particular, maintains lymphoid and B-cell lineage specific programs in readiness for induction. Ikaros and E2A together regulate the expression of flt3, IL7-Rα, TdT, Ebf1, Rag1 and Rag2 in the LMPP cells (Ng et al. 2009; Dias et al. 2008). These genes are not expressed in the LMPP cells of Ikaros null mice (Yoshida et al. 2006). Many of their 5'-regulatory regions contain Ikaros binding sites and the LMPP cells isolated from Ikaros null mice lack B cell potential while exhibiting reduced T cell potential (Ng et al. 2009, Yoshida et al. 2006; Yoshida et al. 2010, Wang et al. 1996).

Ikaros dominant negative mutant mice do not have any lymphoid potential, possibly owing to the lack of compensation of Ikaros function by Helios or Aiolos, members of the Ikaros family that bind the same regulatory sites (Georgopoulos 2002; Georgopoulos et al. 1994). Also the Ikaros null defect is likely to be concentrated in the Common Lymphoid Progenitor (CLP) cells, which make up most of the B cell compartment while leaving the T cell progenitors, the early thymocyte progenitor (ETP) cells, more intact.

The PU.1 negative mice exhibit a similar phenotype to Ikaros null mice with a lack of T, B, NK and myelomonocytic cells and defects in the LMPP compartment (Scott et al. 1994; Scott et al. 1997; Ramirez et al. 2010). However, PU.1 is not strictly required for B cell development because B lineage cells can be grown from PU.1 -/- cells on stromal cell support and in the presence of SCF and IL-7. The most prominent role of PU.1 seems to be to act as a dosage-dependent lineage switch, since high levels of PU.1 in fetal liver progenitors generate macrophages and lower levels of PU.1 are required for B lineage specification (DeKoter & Singh 2000, Laslo et al. 2008).

The E2A gene is in the same HLH family as EBF1 and comes in two forms, E47 and E12. The transcription factors E2A and Foxo1 are both required for the generation of the B lymphoid system (Bain et al. 1994; Amin & Schlissel 2008; Dengler et al. 2008). Recent genome-wide ChIP-seq studies (Lin et al. 2010; Rothenberg 2010) have placed Foxo1, together with E2A, firmly as co-regulators of Rag1/2 as well as IL7Rα. They

are present together and required for the initiation of EBF1 expression. It would be interesting to see the overlap between Ikaros and PU.1 bound cis-regulatory sites with E2A/Foxo1 bound sites (Yoshida et al. 2010). It seems that Ikaros is involved in earlier phases of development and prepares the ground for E2A and Foxo1 to activate EBF1 expression. E2A and Foxo1 demonstrate the power of combinatorial actions of transcription factors, since neither are B-cell specific on their own. Their interactions are demonstrated at the genetic level by haploid co-insufficiency: heterozygous knockouts of the two factors have a blockage in B cell development whereas neither on its own has such a blockade (Lin et al. 2010).

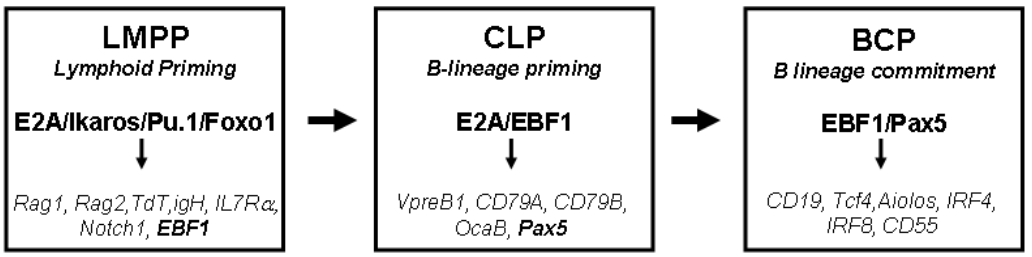

Figure 1. Transcription factors crucial for each stage of early B cell development and a set of relevant targets. EBF1 and Pax5 are highlighted as targets since they continue the cascade to the next stage (Bryder & Sigvardsson 2010). LMPP: lymphoid myeloid primed progenitor, CLP: common lymphoid progenitor and BCP: B lymphoid committed progenitor.

2.1.3. B lineage priming and commitment

B cells express several transcription factors throughout the B cell developmental cascade that are important for their functions and their gene expression program. These include Ikaros, PU.1, E2A, EBF, Pax5, LEF-1, SOX4, IRF4, IRF8, OBF1, Aiolos as well as other factors that are either classified as constitutive transcription factors or present only in very specific stages of B cell development (Ramírez et al. 2010).

The elimination of E2A and EBF1 leads to a complete block of B cell development at the pro-B-cell stage prior to D to J_H recombination of the IgH locus (Bain et al. 1994; Lin & Grosschedl 1995). These transcription factors thus play key roles at the onset of B-cell differentiation or specification, and their expression alone with the Rag1/2 genes is sufficient for D to J_H recombination (O'Riordan & Grosschedl 1999; Romanow et al. 2000). The co-operation of E2A and EBF1 is demonstrated at the functional and genetic level by the synergistic effect that the double heterozygote knockout has on several B-cell specific genes including λ5, Rag1/2, mb1 and Pax5 (Figure 1) (Lin et al. 2010).

E2A and EBF1 together are required for the priming of the B cell program and the expression of the Pax5 gene (O'Riordan & Grosschedl 1999; Bryder & Sigvardsson 2010) (Figure 1), with involvement from PU.1, IRF4 and IRF8 (Decker et al. 2009). The switch from E2A/EBF1 dependent expression of B cell genes to EBF1 and Pax5 dependent expression is typified by the switch of regulation of the EBF1 gene expression by E2A and STAT5 (via IL7) to the regulation of the EBF1 gene in a positive feedback loop by Pax5 (Roessler et al. 2007) (Figure 1).

2.1.4. The role of the Pax-5 transcription factor in the B cell program

E2A, EBF and Pax5 are present throughout B cell development. Pax5, increasingly recognized to act together with EBF1 (Treiber et al. 2010), is considered the lynchpin of the B cell transcription factor network (Figures 1 and 2). While E2A and EBF1 act mainly to induce B cell genes, Pax5 is required for irreversible B cell commitment (Nutt et al. 1999b; Mikkola et al. 2002; Rolink 1999; Cobaleda et al. 2007b).

The role of Pax5 as the arbiter of B cell fate was first seen in cells from Pax5 knockout mice (Nutt et al. 1999b; Mikkola et al. 2002; Rolink 1999). Pax5 -/- mice are arrested at the pro-B-cell stage of B cell development, after undergoing the D to J_H rearrangement and the proximal V_H to DJ_H rearrangement (Nutt et al. 1997). They fail to undergo the distal V_H to DJ_H rearrangement or express a functional pre-B-cell receptor. The lack of commitment of the Pax5 knockout cells to the B-cell lineage is seen in that they are able to be de-differentiated in culture with IL-7 and behave like STRC stem cells (Nutt et al. 1999b; Rolink et al. 1999). They are able to reconstitute the entire bone marrow compartment, except the B cells, under appropriate culture conditions as well as *in vivo* after transplantation into irradiated mice. A similar phenotype is seen in E2A deficient bone marrow derived cells, which are able to assume similarly diverse lineage fates as the Pax5 deficient cells. This suggests that progenitor cells exhibit lineage plasticity before the activation of the B lineage program including EBF, E2A and finally Pax5. Pax5, as the latest factor in this cascade, which still causes plasticity after removal, can then be termed the B cell commitment factor (Cobaleda et al. 2007b).

Lineage regulators are seen to carry out their tasks chiefly through suppression of alternate lineage choices as well as by activating lineage appropriate genes. In contrast to the B-cell specification factors E2A and EBF, Pax5 also represses alternative gene expression programs such as those regulated by Notch1, which is required for T cell commitment (Figure 2) (Rolink et al. 1999). Pax5 is also able to positively regulate its own expression and is expressed and required throughout the B cell program (Mikkola et al. 2002; Decker et al. 2009; Roessler et al. 2007) (Figures 1 and 2).

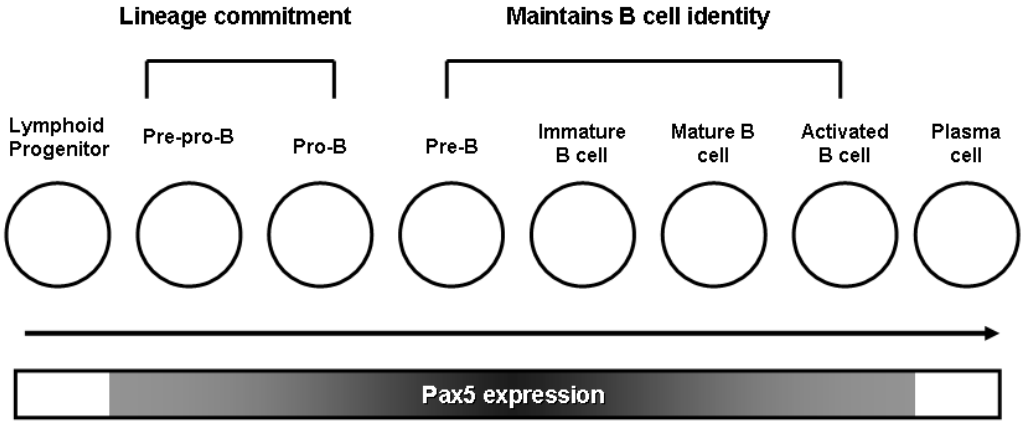

Figure 2. The expression and roles of the Pax5 transcription factor during B cell development.

2.1.5. Avian B cell development and the DT40 cell line

Removal of the chicken bursa of Fabricius was first used to demonstrate that bursa-derived cells, termed B cells, are the source of antibody responses (Cooper et al. 1965). Instead of the foetal liver and bone marrow, avian species utilize gut-associated lymphoid tissues (GALT) for primary B-cell development (Veistinen & Lassila 2005; Weill et al. 2004). Avians also differ from human and mouse in that they have a very simple immunoglobulin (Ig) locus. They make use of homologous recombination (called gene conversion) to transfer genetic material from unexpressed immunoglobulin variable (V) region pseudo-genes in the vicinity of the rearranged active Ig locus (Arakawa H et al. 2004), resulting in Ig diversity. Mammalian species, such as rabbits and sheep, also carry out part of their B-cell development in GALT tissues. For instance, rabbit and swine use gene conversion to generate diversity, whereas sheep rely mainly on somatic hypermutation. The chicken bursa of Fabricius is however the most highly specialized primary B-cell organ.

Chicken B-cell development can be divided into three stages called pre-bursal, bursal and post-bursal (Figure 3) (Toivanen & Toivanen 1973). Both the Ig heavy and light chain variable genes are rearranged by embryonic day 5 (ED 5), and prior to the entry into the bursa (Mansikka et al. 1990). Avian species lack surrogate light chain genes and their B cells enter the bursa with a fully formed, if undiversified, IgM receptor (Figure 2). After embryonic day 15 (ED 15), Ig diversification proceeds via gene conversion and lasts until 4–6 months after hatching. B cells in the bursa are located in the follicles where they proliferate rapidly, and over 95% of the bursal B cells are deleted through apoptosis (Weill et al. 2004; Veistinen & Lassila 2005). The bursa contains stem cells that are able to reconstitute the B-cell compartment and restore the morphology of the bursa after cell ablation. Post-bursal B cells emigrate from the bursa just prior to hatching. The lifetime chicken B-cell repertoire is generated in the bursa which involutes after 6 months. The B cell repertoire is further diversified via somatic hypermutation in the germinal centres (Figure 3) after B cell activation. Avian species also have a dedicated plasma cell organ called the Harderian gland that contains Ig-secreting terminally differentiated plasma cells and is located around the orbit of the eye (Mansikka et al. 1989).

Despite the differences, B-cell development in avian species and in mammals is a very similar process (Figure 3). The similarities are even greater at the molecular level and at the level of the regulatory networks (Weill et al. 2004, Koskela et al. 2003; Wu et al. 2004). Comparisons among species reveal the evolutionary plasticity inherent in the regulatory networks. Similar mechanisms are modified to produce different outcomes (Arakawa et al. 2004, Article IV). This plasticity is also employed at the single organism level to recycle genetic components, and its study in model organisms is likely to help to unravel the way human systems are regulated.

The chicken DT40 cell line model offers an ideal system for studying genes that are important for B cell function (Brown 2003; Alinikula et al. 2006; Smith et al. 2004; Wahl et al. 2004). DT40 is an avian leukosis virus induced B cell line of bursal origin that is constantly undergoing immunoglobulin light chain gene conversion at a high

rate (Buerstedde et al. 1991). Upon transfection, sequences are rapidly integrated into the DT40 genome via homologous recombination. The results obtained are also applicable, at least to some extent, to B cell lymphomas and leukaemia in man (I; III; Delogu et al. 2006; Kurosaki 2002). Furthermore, the DT40 system is well suited for studying B cell signalling and apoptosis (Kurosaki 2002). B cell signalling can be studied for instance via phosphorus incorporation after the signal is activated, or alternatively calcium mobilization can be assayed.

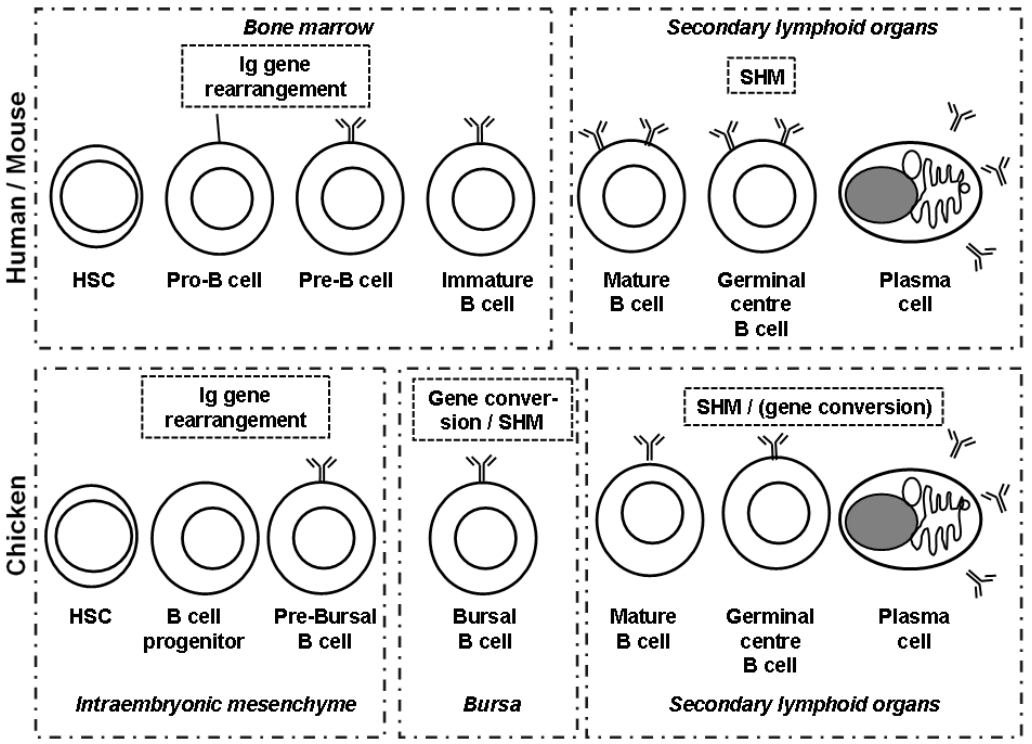

Figure 3. Comparison of chicken B cell development with human and mouse B cell development.

2.1.6. Germinal centre B cells and the BCL6 protein

The survival of plasma cell precursors, the germinal centre B cells, is dependent on the BCL6 transcription factor, which among other effects counteracts the expression of p53 (Phan & Dalla-Favera 2004) and upregulates the Bcl2-like anti-apoptotic protein Bcl-X_L (Klein & Dalla-Favera 2008). The B cell survival factor BAFF also provides an important survival signal while the cell population in germinal centres increases rapidly (Schneider et al. 1999**).**

2.1.7. The B cell to plasma cell transition and plasma cell maturation

B cell activation, usually with the help of cognate T cells, leads to the emergence of memory B cells as well as plasma cells (Figure 4). The steps in B-cell activation

leading to affinity maturation to produce high-affinity antibodies take place in the germinal centres.

Plasma cells are the antibody factories of the organism and thus the effector cells of the humoral immune response (Janeway et al. 2008; Klein & Dalla-Favera 2008). Plasma cells have an enlarged endoplasmic reticulum (ER) as a sign of their high secretory activity. Their ability to secrete large amounts of antibody depends on the activation of the unfolded protein response (UPR) (Todd et al. 2008).

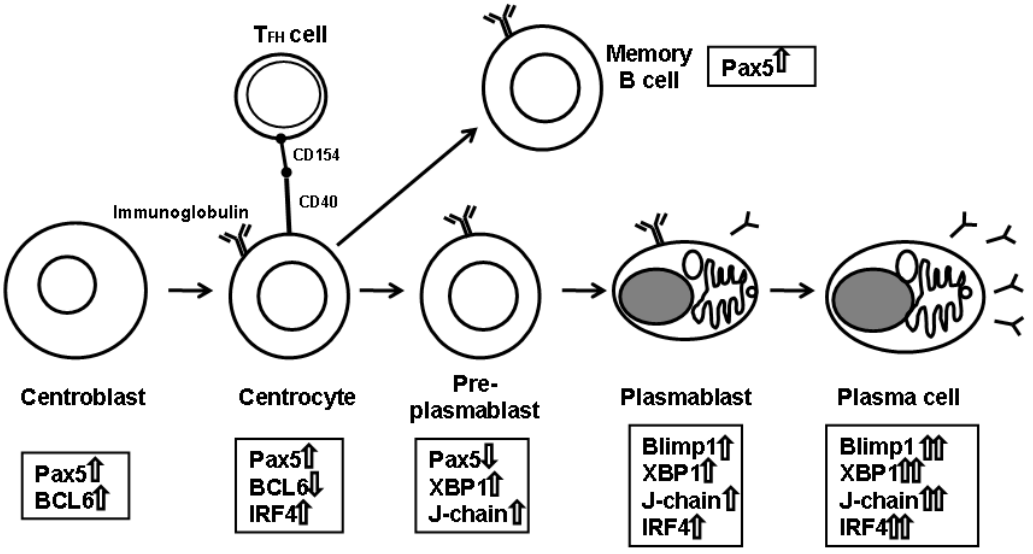

Figure 4. The expression level changes of the key regulators at various stages of B cell to plasma cell transition. T_{FH} (follicular helper T cells)

Since mature terminally differentiated plasma cells produce high levels of secreted immunoglobulin, the unfolded protein response (UPR) is induced and required for the later steps of plasma cell maturation (Todd et al. 2008). The XBP1 protein plays an important role in enabling the full induction of UPR. ER stress leads to the activation of IRE1 that splices, in the cytoplasm, the XBP1 transcription factor messenger RNA to produce the active form of the protein that promotes Ig secretion by upregulating the secretory J-chain of the immunoglobulin gene and expanding the secretory apparatus (Todd et al. 2008; Shaffer et al. 2004).

The key mechanisms by which B-cell activation produces plasma cells are still not completely known (Calame 2008). The Blimp-1 protein is required for the development of immunoglobulin-secreting cells and for maintenance of long-lived plasma cells (LLPCs) (Martins & Calame 2008). Secondary signals, which usually come from the cognate follicular helper T cells, play an important role in shifting the balance of the regulatory network towards the plasma cell fate. The regulatory logic of the B cell to plasma cell network and the specifics of the transition will be explored further in the Discussion (6.1.2 – 6.1.4). The network of transcriptional regulators promoting the B cell phenotype includes at least Pax5, BCL6, MITF and IRF8. Blimp-1, IRF4, XBP-1 promote the plasma cell fate (Klein & Dalla-Favera 2008; Calame

2008; Schmidlin et al. 2009) (Figure 4). The IRF4 gene is upregulated prior to the Blimp1 gene, and Pax5 and BCL6 need to be inactivated before the plasma cells develop (Figure 4) (Klein & Dalla-Favera 2008; Lin et al. 2002).

2.2. Bioinformatics of gene expression data analysis

2.2.1. A multidimensional view of genome annotation

Systems biology recognizes biology as a data-rich science and stresses the importance of integrating data from disparate sources to answer biological questions (Aderem 2005; Aerts et al. 2006; Joyce & Palsson 2006). Systems biology aspires to be a synthetic science that aims to put the pieces back together in order to test our understanding of the higher level phenomena that do not hinge on the actions of single genes or proteins, emphasizing the relationships between them and interactions at the different hierarchical levels of the biological systems. Hierarchical levels mean a level of representational detail that can be tackled experimentally, independently of other levels. Different levels interact with each other and may overlap.

In the post-genomic era (International Chicken Genome Sequencing Consortium 2004; Burt 2005; Furlong 2005; Lander et al. 2001; Venter et al. 2001) ideas about multidimensional genome annotation usually highlight the difference between knowing which parts (genes, RNAs, proteins) are present inside the cell at any given time as opposed to what they do and how they interact to carry out biological processes (Reed et al. 2006; Joyce & Palsson 2006; Brent 2008). One needs to go from profiling and describing parts-lists to reconstructing the regulatory networks (Karlebach & Shamir 2008). To do this one needs to perform, often in high-throughput, functional perturbations that directly probe the functional significance of those parts (Erfle et al. 2007; Hoheisel 2006; Wheeler et al. 2005) and aim to find out how the concentrations or activities of the network components change during the response (Schadt 2009).

2.2.2. Role of nucleic acid quantification in biological investigations

Arguably the measurement of RNA and especially mRNA is, to date, the most successful way to tackle the goals of systems biology to investigate how genes as a group generate biological processes and regulatory networks that are active in a cell at a given time.

The ease of measuring nucleic acid concentrations in cells has to do with the chemical uniformity of RNA and DNA strands, which enables their efficient extraction and uniform assay conditions, while each species can be identified to a high degree of accuracy via specific hybridization of the complementary DNA strands. This reflects the primary function of DNA and RNA as information storage and carrier molecules (Watson & Crick 1953). Proteins have a far more varied chemical make-up and cannot yet be analyzed in parallel to the same degree as nucleic acid polymers, even though

mass spectrometry-based methods are slowly bringing proteomics towards the kind of level of throughput that DNA assays have enjoyed for years (Ong et al. 2002).

However, mRNA levels do not always correspond well to protein levels. The levels are related, although the correspondence is often not a direct one due to regulatory mechanisms that affect either the rate of protein translation or the persistence of proteins once they have been made (Waters et al. 2006). The activity of proteins also varies dramatically according to the post-translational modifications they have, some of which can be detected with mass-spectrometry.

Gene expression microarrays were developed to meet the challenge posed by the explosion of new knowledge about genes and genomes generated by EST sequencing projects and whole genome sequencing by enabling the relative quantization of tens of thousands of RNA species at the same time. Gene expression arrays exploit the hybridization of labeled DNA or RNA strands to their complementary strands that are fixed to a solid support (Kapur et al. 2007; Schena et al. 1995).

Microarrays also have limitations. Most arrays have not been designed to detect differences in the transcripts due to alternative splicing or polyadenylation (Kapur et al. 2007). Depending on the probe design these regulatory events could manifest themselves as either a loss or a gain of transcripts, or may go entirely unnoticed. Some gene expression arrays, such as Affymetrix exon arrays, do however address these post-transcriptional regulatory mechanisms to some extent (Kapur et al. 2007).

High-throughput sequencing is addressing the short-comings of microarrays (Wang et al. 2009) by providing direct information of the RNA or DNA content of the cell instead of the proxy information provided by the probe based on microarray technologies (Shendure & Ji 2008). The RNA-seq technology is able to measure alternative splicing and poly-adenylation more effectively than arrays but extracting this information from the sequence data requires a great deal of processing and the results are, with the current technology, more probabilistic than certain (Wang et al. 2009).

2.2.3. Evolution of array platforms for gene expression analysis

The gene expression microarray technologies have undergone many technological changes during the last ten years (Schena et al.1995; Gautier et al. 2004; Kapur et al. 2007). They have matured as a technology from being commonly used to profile a few hundred to a few thousand genes on a nylon filter to profiling entire mammalian genomes simultaneously. There has been a corresponding upsurge in the popularity of the technology as well as a convergence on the platforms and the analysis methods used to turn raw data into biologically meaningful results (Allison et al. 2006).

Three or four basic types of arrays have been commonly used and they can be classified as two-colour or one-color depending on whether two or more differently labelled samples are hybridized together competitively or whether only one sample at a time is hybridized to the array. The arrays are usually only able to measure the ratio of the gene product in the samples under comparison and not the absolute quantity of the

RNA in the cell, although the intensity of the signal is roughly indicative of the absolute analyte quantity.

The oldest array platforms are based on cDNA strands or bacterial colonies (Grunstein et al. 1975) expressing the DNA of interest fixed to a nylon filter support with UV light cross-linking. The probe cDNA is labelled with radioactive phosphorus, usually with P32 or P33, and visualized with a phosphorimager. The technology has several drawbacks such as a low feature density which means that arrays can be as large as 22x22 cm, making them unwieldy to handle. The large size of the array makes it highly likely for some local differences in the hybridization efficiency to occur, and these differences need to be taken into account when dealing with the data (Wu et al. 2003, Edwards 2003). Since the arrays are one-color, replicate array performance should be very uniform in order for the results of two experiments to be comparable.

Glass-slide arrays produced by academic groups (Schena et al. 1995) or government centres are usually made with either synthetic and relatively long 60-mer oligonucleotides or with cDNAs from a cDNA library that are deposited to the slide with metallic print tips. The advantage that these kinds of arrays have over nylon filter arrays is that they have a much higher density and often a larger number of spots as well. Importantly, they are also hybridized in the two-color mode, which helps to offset irregularities in the manufacturing process since the correct ratio of the transcripts from two different samples is always known. Synthetic oligos are preferable since their sequence is completely known and the sequence can be chosen to minimize chances of a cross-reaction with other genes or genome sequences.

In-situ synthesis is another method of depositing oligos onto a substrate. The best known vendor is Affymetrix that uses a high density photolithographic mask and 16-22 oligos of 25-mer length per gene (Pease et al. 1994, Southern 1996). In order to gauge the specificity of the hybridization to the relatively short 25-mer oligonucleotide, a mismatch oligo with one base pair difference was originally used as a control, and the signal was measured as a difference between the intensities of the two oligos (Gautier et al. 2004). More recently, this practice has been abandoned (Kapur et al. 2007).

2.2.4. Design of high throughput gene expression studies

It is important to have a clear goal in mind when designing and performing array studies (Miron & Nadon 2006; Allison et al. 2006). While some open-ended array studies may be done in order to assemble a gene expression atlas across various tissues or developmental stages (Hoffmann et al. 2002, Hoffman et al. 2003; Rhodes et al. 2007; Kilpinen et al. 2008, Shaffer et al. 2006), arrays usually perform best in classifying a phenotype in great detail and can also more readily aid in suggesting a hypothesis for further study (Shaffer et al. 2002; Shaffer et al. 2001).

Microarray studies require sufficient biological and technical replicates to yield valid results, in practice a minimum three to five replicates must be used. Biological replicates are independently obtained biological samples (Allison et al. 2006). Technical replicates measure the same biological sample with replicate arrays and are

thereby mainly useful for array quality control purposes. Since the quality of most commercial array platforms is such that the technical variation is much smaller than the biological variation, technical replicates are generally not needed.

Assays on gene expression that do not include experimental manipulations are often termed descriptive. While they can tell which genes or gene programs are active in specific tissues, they do not usually tell of their importance to that tissue or the developmental process it is undergoing. Gene expression studies on normal tissues are, however, often necessary for understanding diseased tissues or as a control for experimental manipulations. Malignancies, such as B cell leukemia (Shaffer et al. 2006), often retain characteristics of the tissue that they originated from, and it can be very useful to compare their expression profiles to the corresponding primary normal cell or tissue expression profiles.

Highly parallel microarray studies ideally require high throughput experimental techniques so that the global effects on multiple genes can be assayed to form a clearer picture of the regulatory relations and networks. The most effective manipulations often involve gene inactivation or over expression (Shaffer et al. 2006).

Array studies can also assay a time course, which can be especially useful for studying a cellular differentiation process (Elo et al. 2007). The results of the time course can be classified and clustered based on the response such as early or late activating genes. Depending on the type of manipulation, this may or may not indicate that the genes with a similar expression profile are coregulated or belong to a common functional group. When constructing gene regulatory networks, time-course studies have the advantage that causal direction can be deduced, at least in some cases (Zhao et al. 2006, Elo et al. 2007).

2.2.5. Pre-processing and normalization of microarray experiments

Microarray data are subject to multiple sources of variation, of which biological sources are of interest and most others are only confounding (Draghici et al. 2006; Workman et al. 2002). Pre-processing steps of the data take place before data analysis and seek to reduce the relative error between replicates by removing these systematic sources of variation. Since biological sources of variation introduce outlier data points relative to the systematic effects, it is important to utilize robust methods that are less affected by outliers. Most data normalization methods make the assumption that a majority of the genes are not differentially regulated between samples. When utilizing whole-genome arrays this assumption is usually valid.

Pre-processing of microarray data includes image analysis, normalization and data transformations (Allison et al. 2006; Gautier et al. 2004). Image analysis entails the derivation of numerical values belonging to a known probe on the array from the pixel intensities of a microarray image. It is highly technology dependent and is usually performed by software from the array manufacturer or the scanner manufacturer. The image analysis software will produce various values such as the signal intensity of the spot and background and possibly quality control parameters. Various methods for

removing the signal background have been developed (Edwards 2003). However, if the background is not very high (favourable signal to noise ratio), it is better to often just not attempt background correction. Subtracting the background can lead to issues that can complicate the downstream analysis, such as very low or negative signal values. Very low signal values can produce anomalously high fold-change (FC) values, for instance.

In statistics, a transformation is often carried out in order to ensure that the data follows the normal distribution. The most common data transformation in microarray data analysis is the log transformation, usually in base 2. While the log transformation is less intuitive for the scientist examining the results, it has certain benefits. Microarray data has both additive and multiplicative error terms. The log-transform equalizes multiplicative intensity dependent variance by decoupling random multiplicative error from the true signal intensity (Kreil et al. 2005, Huber et al. 2002). The microarray data distributions tend to be asymmetric around the mean with more values having low signal levels. The log-transformation also converts a formerly asymmetric data distribution into a more symmetric and Gaussian-like one. This allows statistical tests such as the t-test and ANOVA, which assume the data is normally distributed, to be used.

At low levels of signal intensity the additive errors dominate, such as the variation in the background signal. Thus, variation in the log-transformed data tends to increase at lower signal levels. The linear logarithmic (lin/log) transformation can help by transforming the data linearly at a specific threshold level and carrying out a log transform otherwise. A generalized log transform such as an asinh transform uses a model to estimate the parameters for the additive and multiplicative error terms from the data (Huber et al. 2002, Kreil et al. 2005) and adjusts the transform to equalize variance as a function of signal intensity. The generalized log transform is approximately equivalent to the log transform at large values and behaves linearly as the intensity values tend towards zero (Durbin & Rocke 2003).

Normalization is the process of reducing statistical error in repeated measured data. Every microarray platform tends to have accepted solutions for data normalization and transformations (Allison et al. 2006). They are accessible from the R/Bioconductor repository, which has become the standard data analysis platform for microarrays (Gentleman et al. 2004; R Development Core Team 2009). The variance stabilizing normalization (VSN) is part of the vsn Bioconductor package (Huber et al. 2002). It carries out an asinh generalized log data transformation and forces the arrays to have the same central tendency with respect to a set of non-changing genes. An iterative least squared fit is used to identify the subset of non-differentially expressed and non-outlier genes. The method assumes that, at most, 50 per cent of the genes are differentially expressed. Thus the VSN method is relatively robust to outliers. A similar method, called variance stabilizing transformation (VST), is used to normalize Illumina array data in the lumi Bioconductor package (Lin et al. 2008).

Some systematic sources of variation in microarray data are non-linear in nature. A method commonly referred to as loess normalization (locally weighted scatter plot

smoothing) is a non-linear method that performs detrending of the signal ratio between control and measurement samples as a function of the local average of the signal intensity (Workman et al. 2002, Bolstad et al. 2003). All the arrays in the experiment can also be forced to follow a reference distribution that can be an average of the original distributions or perhaps even the normal distribution. This is often referred to as the quantile normalization (Workman et al. 2002, Bolstad et al. 2003).

Non-linear normalization methods can be placed on a continuum, ranging from conservative to non-conservative, according to how much between-sample variation in the data distributions they can tolerate. The loess and quantile normalizations are not conservative normalization methods. They can be used if no strong asymmetries are expected, on average, in differential expression between the samples. Otherwise, the quantile normalization in particular can amplify noise and create spurious array results.

Physically large and poorly manufactured arrays, in particular, often benefit from surface detrending, which can be carried out with loess smoothing. Using loess one can estimate whether some areas of the array have a higher signal level, on the average, than others. In general, microarray normalization is well understood. However, new technologies that, for instance, carry out experiments on array platforms often require the return to the normalization methods employed on earlier generations of gene expression platforms (Leivonen et al. 2009).

2.2.6. Statistical inference for microarray experiments

Fold-change (FC) estimates of expression difference between groups of samples can be used for rank ordering of genes (Shi et al. 2008). One major disadvantage of FC estimates, however, is that they do not take the variance of the samples into account. Using the fold change for differential expression does not take into account gene specific biological variance in gene expression, even when global variance stabilizing transformations, such as VST, are applied (Allison et al. 2006; Murie et al. 2009). Since variability in gene expression measurements is partially gene-specific, statistical tests of differential expression are preferred over FC for inference (Allison et al. 2006).

Microarrays typically have small sample sizes, three to five samples being a common occurrence. Small sample sizes make it impossible to estimate the gene-specific variance reliably. Thus, techniques have been developed that are generally referred to as "variance shrinkage", which capitalize on the parallel nature of microarray measurements to borrow information across genes to improve variance estimates and thereby increase the statistical power of the tests. The gene-specific and global variance estimates are weighted differently depending on the methodology used. The significance analysis of microarrays (SAM) was an early method to employ this approach (Tusher et al. 2001). The Bayesian approaches to weighting these two variances are considered most successful to date (Murie et al. 2009). The eBayes statistic from the limma package (Smyth 2004) combines a fitted linear model of gene expression data with a variance estimate into a moderated t-statistic.

The test statistic can be used for ranking the genes or probes on the array based on differential expression. The significance of the gene expression difference is expressed as the p-value which denotes the probability that the expression of the gene follows the null distribution, i.e. is not differentially expressed. However, in order to determine the significance of the results one has to assume that the test statistic follows a certain distribution, usually the t-distribution. Permutation based null distributions or resampling can be used to avoid making assumptions about data distributions, as is the case with the SAM method. The granularity or accuracy of the permutation based distribution is an issue, however, and reduces the power of this approach to detect differential expression. The same is true for non-parametric test statistics, such as the Wilcoxon test, which also have their own biases. Parametric methods are most commonly employed in differential expression analyses.

Gene expression analyses have not only low sample sizes but make thousands or tens-of-thousands of parallel measurements as well. While these measurements can be used to improve variance estimates, they also present a multiple testing problem (Allison et al. 2006, Shaffer 1995, Storey et al. 2003; Clarke et al. 2008). If 20000 measurements are made at the $p<0.05$ significance level then the number of expected false positive results would be 1000. In order to test each null hypothesis “independently" from the outcome of others, one can control the so called family-wise error rate (FWER). This method, termed Bonferroni correction, divides the p-value of the single gene test by the number of tests performed and ensures that the probability of a single false positive determination within the entire experiment matches the specified threshold. However, when the number of measurements is large, this tends to lead to an overly large false negative rate. The False Discovery Rate (FDR) is defined as the rate by which false discoveries occur or the expected proportion of false positives among all significant hypotheses (Benjamini & Hochberg 1995). In practice a Bayesian modification, positive FDR (pFDR), defined as the rate by which discoveries are false, is employed. The pFDR is used as a basis for the estimating q-values to decide on the minimum pFDR over which a statistic can be rejected (Storey 2003; Storey & Tibshirani 2003). Usually multiple testing corrected q-values are used, instead of p-values, to uncover significant changes in gene expression patterns.

Finally, statistical significance does not always denote biological significance. Especially with large numbers of samples small changes can be statistically significant but are very unlikely to exert a biological effect. It can be helpful to prioritize the differentially-expressed genes with the highest FC, as recommended by the MicroArray Quality Control (MAQC) project (Shi et al. 2008). An FC of around 2 is often considered likely to be biologically relevant. However, low-throughput experiments are needed to validate and further investigate the hypotheses generated on the basis of high-throughput data.

2.2.7. Gene class testing as a tool for gene expression analysis

The biological significance of findings can be very difficult to determine from a list of differentially-expressed genes. Microarray results, however, usually consist of differential expression of individual genes or predefined groups of genes. These

groups of genes can be genes belonging to the same biological pathway or sharing a similar function or be co-expressed in a different study (Eisen et al. 1998). Co-expressed genes can also be regulated by the same transcriptional regulators. Therefore, it is advantageous to compare the results from one´s own study to other biological experiments or to gene sets related to a specific biological function or process. This is termed gene class testing (GCT) (Nam & Kim 2008; Allison et al. 2006).

Gene Ontology (GO) is an attempt to systematically unify and describe biology in a computer readable manner (Ashburner et al. 2000; Gene Ontology Consortium, 2010). It takes the form of a directed acyclic graph (DAG) in which terms are nodes and relationships among them are edges. It is divided into three independent ontologies: molecular function (MF), biological process (BP) and cellular component (CC). Genes can be associated with the GO terms.

Other projects such as KEGG (Kanehisa & Goto 2000; Kanehisa et al. 2008) and Biocarta (www.biocarta.com) also describe pathways and connect genes to them. Especially the KEGG is developing rapidly (Kanehisa et al. 2008). The Molecular Signatures Database (MSigDB) contains canonical pathways from Biocarta and KEGG, genes up or down regulated in various biological experiments (Subramanian et al. 2005) as well as predicted targets of miRNAs and transcription factors (Xie et al. 2005).

Methods for GCT can be divided into two categories: individual gene analysis (IGA) and gene set analysis (GSA) (Nam & Kim 2008). The differences are illustrated below (Figure 5). In brief, IGA determines individually differentially expressed genes and then looks at the gene lists that satisfy a certain cut-off threshold for overrepresented biological themes. These lists are then tested with the Fisher's exact test or a similar statistic to determine enrichment probabilities, which are then corrected for multiple testing using same procedures as in the microarray differential gene expression analysis. GSA, in contrast, looks at differential expression at the level of the sets directly and is not dependent on cut-offs for differential gene expression. A scoring metric is used for the tendency of the set to lie in either extreme of the ranked list of expression values. The p-values for the differential expression of the gene set are determined either by assuming the scores are distributed in a certain way (parametric means) or by permutation of the sample labels, the gene labels or both. Multiple testing corrections are applied to the p-values.

The gene set enrichment analysis (GSEA) tool was perhaps the first GSA tool (Subramanian et al. 2005). Others such as the Gene Set Analysis (GSA) tool (Efron & Tibshirani 2006) have been developed. The list based IGA tools, such as DAVID (Huang et al. 2009; Hosack et al. 2003), are still being used widely, perhaps due to the easy web access. Tools of the GSA type are more demanding to implement on the web, since they require submission of the raw expression data, and permutation based p-value determination requires longer computational times.

However, the results of IGA, especially, vary widely depending on the methods used for differential gene expression and the cut-offs that are chosen (Nam & Kim 2008;

Rhee 2008). The various methods for carrying out GSA tend to give more consistent results (Liu et al. 2007). Therefore, GSA methods are preferred for data reduction and for gene class testing in the analysis of microarray data.

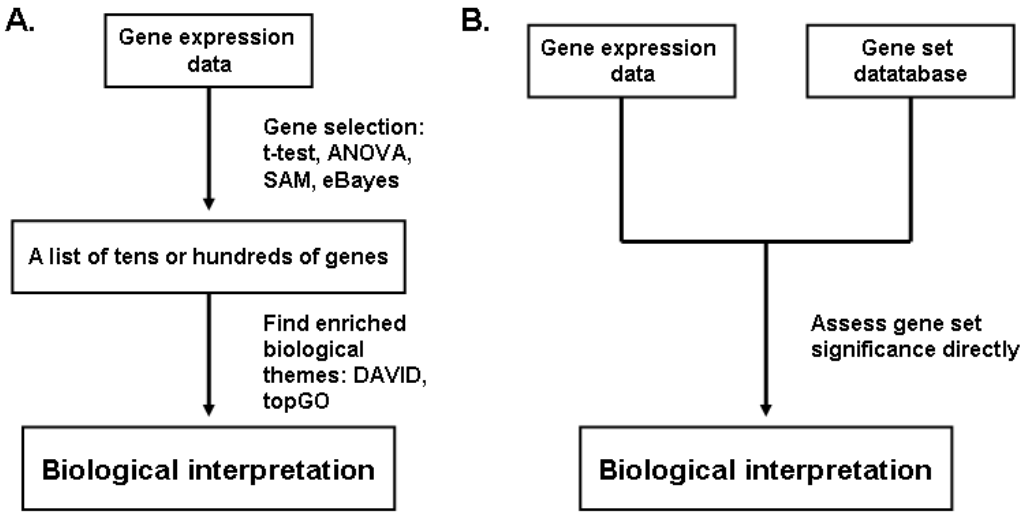

Figure 5. Comparing gene set analysis (GSA) with individual gene analysis (IGA). A) Results of the individual gene analysis depend on the method of gene selection. B) Gene set analysis is direct and produces more reproducible results (Nam & Kim 2008).

2.2.8. Meta-analysis to combine diverse studies

Meta-analysis approaches can compare diverse datasets and weigh their contribution to the test statistic according to their biological significance or the quality of the experimental data (Troyanskaya 2005). Many statistical tools have been developed for gene expression meta-analysis. The R/Bioconductor-based RankProd package (Hong et al. 2006) uses rank products to combine ranks across studies and estimates p-values with permutation based methods. Multiple testing correction and visualization of the statistical significance are also implemented.

2.2.9. Standards and public availability of microarray data

The public availability of large sets of microarray data is changing the way results of experiments are validated and hypotheses are generated. The requirement of submitting all published gene expression data to public repositories such as ArrayExpress (Parkinson et al. 2009) or GEO (Barrett et al. 2009) and the MIAME standard of microarray data submission (Brazma et al. 2001) have been instrumental in this change. Methods of using this data are developing as well (Parkinson et al. 2009). For instance, ArrayExpress and EBI, in the form of ExpressionProfiler (Kapushesky et al. 2004) and Atlas (Parkinson et al. 2009), have tools that not only facilitate storage of the gene expression data but also gene-specific retrieval by experimental details, as well as programmatic access to the data.

Secondary databases that contain well-annotated and integrated expression data sets, as well as tools to analyze them, are playing an important part in making the expression data more easily accessible. The Oncomine database (Rhodes et al. 2007) has, at the moment, over 27000 cancer-related array experiments, where information on the expression of individual genes can be queried across studies. GeneSapiens (Kilpinen et al. 2008) is a database that makes use of a novel normalization enabling direct comparison of gene expression values across different Affymetrix array generations (Autio et al. 2009).

2.3. Perspectives to Network Biology

In order to understand better the regulation of developmental decisions, it can be helpful to study regulatory factors at a higher level of abstraction – that of a regulatory network (Karlebach & Shamir 2008; Barabási & Oltvai 2004; Alon 2007; Barabási 2009; Kim 2009; Schadt 2009).

It is fruitful to treat the cell as an information processing engine. Similarly to a microprocessor, it receives inputs from outside and processes the information in those inputs using its regulatory and protein interaction networks, according to its cell-type specific program (Schadt 2009). The expressed proteins and other components, such as various classes of RNA, could be thought of as memory, which has varying degrees of volatility according to half-life. The program maintains itself via auto-regulatory loops, which often involve transcription factors (Rothenberg 2007; Alon 2007) and epigenetic mechanisms. The most persistent part of the program is largely loaded onto and passed on from one cell generation to another via epigenetic marks on the chromatin (Cairns 2009). The epigenetic marks, as well as the complement of transcription factors that read them, largely define the cell type. The hard-coded circuitry, which is almost unchanging at the timescales of the individual cell or the organism, is the genome (International Chicken Genome Sequencing Consortium 2004; Venter et al. 2001; Lander et al. 2001).

2.3.1. Emerging network biology

Attempts to unify biological sciences with information sciences are termed network biology, integrative biology or systems biology (Hood et al. 2004; Ideker et al. 2001; Kirschner 2005; Liu 2005). They necessitate setting up interdisciplinary networks of people working towards common goals (Liu 2009; Henney et al. 2008).

An essential property in network biology is that biological networks are so-called scale free networks (Barabási 2009) whereby certain nodes, called hubs, are more important for the functioning of the network than others (Barabási & Oltvai 2004; Yu et al. 2007); that certain regulatory motifs occur more frequently in biological networks than one would expect by chance (Alon 2007; Rothenberg 2007); that networks are used to visualize and analyze biological information (Cline et al. 2007; Schadt 2009); and that mathematical modelling can be useful for studying the dynamics and responses of

networks (Kim 2009; Karlebach & Shamir 2008). Immunology has played a pioneering role in the development of the ideas about genetic regulatory networks (Rothenberg 2007). It is anticipated to be important for drug discovery as well (Hopkins 2008; Kramer & Cohen 2004).

2.3.2. Transcription factors in network context

Almost 50 years ago Monod and Jacob described a regulatory circuit they termed double-negative circuit (Monod & Jacob 1961; Staudt 2004). In their example, two enzyme pathways each yield an end-product that inhibits the activity of the other pathway. Double-negative circuits have the property that a perturbation of one of the regulators, even transiently, can push the system towards one of the two cellular states; they are bi-stable (Staudt 2004; Rothenberg 2007; Alon 2007). Monod and Jacob also noted that this kind of a circuit seems ideally suited for developmental decision making switches in animal cells, although they could not name examples at the time.

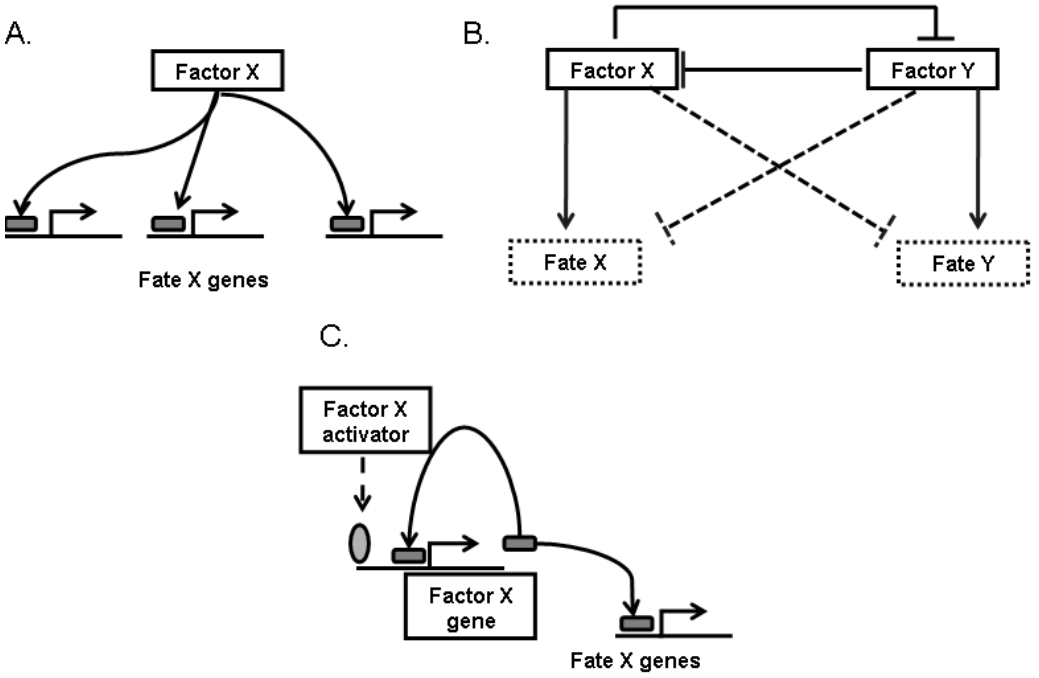

Figure 6. Common regulatory network motifs. A.) Direct positive regulation of a cell type gene battery B.) Balanced mutual repression (developmental toggle switch). C.) Lineage specification factor exhibiting positive auto-regulation. (Rothenberg 2007).

The idea that one transcription factor controls the expression of an entire battery of genes is generally untrue for complex multicellular organisms, as is the idea that cell fate decisions are down to one "master regulator" that can turn any cell into any other cell (Rothenberg, 2007; Komili & Silver 2008). Hardly any gene in the genome has a cell type or tissue-specific expression pattern, and those that do, are usually not regulatory genes, such as the Prostate Specific Antigen (PSA). Introduction of a single factor can change the gene expression program drastically only when the other

components are already present (Figure 6a). Thus, almost all regulatory genes are expressed somewhere else as well, and the genes they induce or repress depend on the context in which they act by virtue of the combination of regulators present or the chromatin regions that have been rendered accessible by chromatin modulators. It can be argued, however, that the active complement of transcriptional regulators largely defines the state of the cell at any given time.

The above reasoning would suggest that all transcription factors are equally important. However, some transcription factors have a preferential role in the transcription factor network that controls cell fate. These lineage regulators, or hubs, can also promote lineage choice, given a cellular state not too far away from the one they normally work in (Figure 6c). Thus, gain of function experiments can drive some cells to adopt a different fate by inducing other transcription factors needed for the lineage specific transcriptional program. They also tend to be constantly expressed during the lineage development (Rothenberg 2007; Ramirez et al. 2010; Bryder & Sigvardsson 2010). In order to do this they rely on epigenetic mechanisms to remain active through cell divisions. They also often exert a positive feedback control on themselves to maintain their own expression (Figure 6c).

Regulatory networks that control cell lineage differentiation frequently exhibit common network motifs. Mutually exclusive differentiation programs often negatively regulate each other (Figure 6b). Auto-regulatory loops make the regulatory programs robust to change, but specific stimuli can toggle the switch from one state to another (Figure 6b and 6c). Stimuli that shift the program are often integrated at the level of the expression or activity of the lineage regulators (Figure 6a). Transition of cells from one state or fate to another can be described as shifts in the levels of these critical regulators (Rothenberg 2007). However, the most important aspect of lineage regulators is often to suppress inappropriate lineage choices rather than promote appropriate ones (Figure 6b) (Ramirez et al. 2010; Bryder & Sigvardsson 2010; Orkin 2000). This makes sense for complex multicellular organisms that need to avoid chaos from too many regulators being active at the same time.

Pax5 is certainly a hub protein in B cells, and a lineage regulator (Figure 6a). The mutual repression between the BCL6-MTA3 proteins and the Blimp1 proteins during the B cell to plasma cell transition is one of the examples that have helped to make influential the concepts of regulatory loops and developmental switches in immunology (Staudt 2004, Fujita et al. 2004; cf. also Figure 6b, 6c and Discussion 6.1.4).

2.4. Evolutionary inference, bioinformatics and whole genome duplications

All biological organisms share a genetic code and evolution may be the greatest unifying principle in biology. The emergence of extant organisms through natural selection and from previously existing organisms undoubtedly constrains the type of

structures and regulatory mechanism that can emerge, as does the process of development itself (Darwin 1872; Erwin & Davidson 2009).

On the other hand, due to its historical and contingent nature, evolution does not easily lend itself to be used as a predictive tool. However, reconstructions of evolutionary past, also termed phylogenetics (Graur & Li 2000) can be helpful in attempting to answer the "why" questions in biology – as well as the mechanistic "how" questions (Mayr 2004).

2.4.1. Sequence alignments

For evolutionary comparisons the sequences are first aligned. Either protein or DNA sequences can be used in the alignments. Even if DNA sequences are used, the alignments are still often made with the help of the protein sequences – especially if the evolutionary distances separating the sequences are substantial. DNA sequences contain information in the third codon position that is lost, in the case of the redundant codons, when nucleotide sequences are translated into protein sequences. On the other hand, the protein similarity matrices, such as the PAM matrix, contain a great deal of information on the chemical and evolutionary similarity of amino acid residues that is not usually taken into account when evolutionary studies are carried out using just the DNA sequences. Consequently, it is usually better to use protein sequence alignments when the evolutionary distances are great since the neutral substitution rate would anyway exceed one substitution per base and render the third codon position information useless.

Alignment programs, such as ClustalW, align nucleotide and protein sequences by placing same or similar nucleotides or amino acids in a column and adding spaces to optimize the alignment where it looks like there has been an indel event (Thompson et al. 1994). An indel is gap in a sequence alignment introduced to account for an insertion or deletion in one or more of the sequences. Distances between the sequences are calculated using a similarity matrix. With the advent of large amounts of sequence information new distance matrices have been computed (Muller et al. 2000).

2.4.2. Methods used to infer phylogenetic relationships between proteins

Phylogenetic sequence analysis can be carried out by at least three different classes of methods. These include the distance based methods, parsimony based methods and maximum likelihood based methods. The distance based methods are also widely used in other applications, such clustering of genes, based on their expression profiles or on other attributes (Eisen et al. 1998).

A phylogenetic tree is then constructed based on the distance matrix. The number of possible trees soon makes it impossible to perform an exhaustive search for even with a limited number of taxonomical units (OTU). Therefore heuristic approaches are used. Perhaps the simplest, and most scalable, method is the UPGMA whereby OTUs are joined in the order of decreasing similarity. A variation called neighbour joining method is usually used in phylogenetics. This method also seeks to minimize the total

length of the branches in the tree and thus produces more accurate trees. Trees can be rooted or unrooted. The root is defined as the most recent common ancestor of all the taxonomic units under study. In the case of molecular evolution an outgroup protein is chosen that is not an ortholog of any of the sequences under study and is presumed to be equally related to every sequence. Choosing the outgroup protein can be difficult because if the sequence is too distant then including it can reduce the quality of the alignment and thus the information content of the distance matrix.

2.4.3. Assessing the reliability of phylogenetic trees

A technique called bootstrap is frequently used to assess the reliability of phylogenetic trees. A bootstrap is a computational technique for estimating a statistic for which the underlying distribution is unknown or difficult to derive analytically (Efron 1982). Using the bootstrap technique typically entails the generation of 1000-10000 pseudosamples of the data by resampling the sequence alignment with replacement to create variation. The tree-building process is then repeated for the pseudosamples and the trees are compared to each other. Nodes that do not receive sufficient support can be collapsed thus generating a multifurcating tree in the place of the original bifurcating tree that has every node intact. There is debate as to how much support a node has to have in order to be considered reliable. A 95% support value is considered reliable but it is unclear as to what the lower values mean or what the cut-off should be (Efron B et al. 1996).

Maximum likelihood methods can be also used and may be more reliable than the best available distance methods. However, they are slow to compute and experimenters often use the simpler methods if they generate sufficiently well resolved trees on their own. Maximum likelihood methods should be, however, used more frequently – especially since computational time is not so much of an issue as of late.

2.4.4. Whole Genome Duplication hypothesis of Susumo Ohno

Nearly forty years ago Susumo Ohno proposed that one or two whole genome duplications (WGD) took place near the origin of vertebrates (Ohno 1970). Recent data from genome wide sequencing projects and paralogous genome segments has confirmed this hypothesis to the extent that it is today considered to be virtually proven (Kasahara 2007; Van de Peer et al. 2010). Since WGD affects organism's every gene simultaneously, it generates large amounts of genetic raw material that can potentially acquire new functions (Van de Peer et al. 2009). For instance, entire regulatory pathways are copied intact and can be co-opted for different roles. Also gene dosage is maintained and genes that are more dosage sensitive, such as regulatory genes, are considered more likely to be retained after whole genome duplication than after regular duplication events. To underscore these differences the genes that are retained after the whole genome duplication are sometimes referred to as ohnologs in the honour of Ohno's contributions. The WGD studies may be relevant to the understanding of gene regulatory networks, since it is expected that such events helped to shape many of them (Erwin 2009). Understanding the history of a gene regulatory network (GRN) may be

helpful in making sense of its present day idiosyncrasies as well as inferring their general design principles.

2.5. Evolutionary systems biology of B-cell differentiation

Despite the B cell development being one of the best studied differentiation processes open questions still remain concerning the mechanisms of gene regulation. The individual factors have been studied but factors do not work in isolation but together in a context-dependent manner.

An evolutionary perspective has proven to be valuable in the past for uncovering the functionally most important subsets among disparate biological details. The conservation of protein sequences across species has been well studied but the study of the conservation of gene expression or gene regulatory networks could form a basis for evolutionary systems biology (Hoffman et al. 2007; Medina 2005).

The chicken DT40 B cell line offers a model system to undertake studies of genetic regulatory networks of avian B cell differentiation. This study looks at the regulatory networks around the Pax5 B-cell transcription factor. Pax5, the key driver of B-cell differentiation, is studied both from a mechanistic point of view as well as from an evolutionary perspective. Tools are developed for characterizing the model system, focusing on B cell development. Analysis of the Pax5 knock-out phenotype is undertaken with gene expression arrays and a cross-species comparative analysis is carried out to determine shared and evolutionarily conserved expression changes at the gene and pathway levels.

3. AIMS OF THE STUDY

1. Development of the avian gene expression array platforms and analysis methods
2. Investigating the avian B cell gene expression program.
3. Studying the gene regulatory network of the avian B cell to plasma cell transition.
4. Characterization of the avian Helios gene and the evolution of the Ikaros family.
5. Cross-species meta-analysis of the Pax5 transcription factor mediated gene regulation

4. MATERIALS AND METHODS

4.1. Array protocols

4.1.1. Making of arrays (I)

Two array platforms were made for the study of gene expression patterns in avian B cells: the B cell EST array (Koskela et al. 2003) and Bursal EST array (I). They contain 288 and 14592 clones, respectively, that were spotted in duplicate.

The Bursa EST library (Abdrakhmanov et al. 2000) was made from cDNA clones extracted from the bursas of 2-week-old chicken of the inbred CB strain. The chicken bursa consists mainly of B cells but Ficoll gradient centrifugation was used to remove contaminating epithelial and red blood cells in order to achieve a B cell content of about 95% B cells. Poly-dT reverse transcribed mRNA was ligated to a linker and cloned into the pSPORT1 plasmid. The average insert length was about 1.3 kb. The library is housed at RZPD (Deutsches Ressourcenzentrum für Genomforschung, Heidelberg, Germany). Sequences were submitted in GenBank under accession AJ392050-AJ3994559. The library is not normalized.

The clones to be spotted onto the smaller B cell EST array were selected using a library of B-cell enriched probes obtained by PCR select subtractive hybridization (Koskela et al. 2003). The probes were hybridized to colony filters containing the dkfz426 bursal EST library. A set of 233 sequences from the dkfz426 library as well as control probes were selected for spotting. The colonies selected in this way were sequenced and the sequences annotated using batch BLAST searches (Altschul et al. 1997). The probes for the array were selected from those sequences based on literature searches.

In order to obtain RNA for the suppressive subtractive hybridization, total RNA of MACS-sorted ChB6a+ Bursal B cells and TCR1+ as well as TCR2+ T cells was isolated with TRIZOL according to the manufacturer's instructions (Life Technologies, Grand Island, N.Y., USA). TCR1+ and TCR2+ T cells correspond to the $\alpha\beta$ and $\gamma\delta$ T subpopulations, respectively. Poly-A+ mRNA was isolated from total RNA with an Oligotex mRNA kit (Qiagen, Santa Clarita, CA, USA). Due to limited amounts of the cell sorted starting material double-stranded cDNA was prepared from poly (A) + mRNA using the SMART system (Clontech, Palo Alto, CA, USA). SMART-generated double-stranded cDNA was digested with RsaI to create smaller blunt-ended fragments that were used as tester or driver in suppression subtractive hybridization that was performed with the PCR-Select kit (Clontech, Palo Alto, CA, USA). The digested tester cDNA from ChB6a+ cells, but not the driver cDNA from pooled TCR1+ and TCR2+ cells, was ligated with two adaptors. The suppression subtractive hybridization and the amplification of differentially expressed cDNAs were performed according to the manufacturer's instructions (Clontech, Palo Alto, CA, USA). The driver TCR1+/TCR2+ cDNA was present in 500-fold excess. The amplified polymerase

chain reaction (PCR) product from suppression subtractive hybridization was hybridized to dkfz426 library filters containing 60,000 bursal cDNA clones, which have an average insert length of 1.3 kb DKFZ (Deutsches Krebsforchungszentrum, Heidelberg, Germany). The probe, enriched for bursal B-cell-specific genes, recognized about 2000 clones, which were sequenced until no more novel sequences were being obtained, that is until saturation levels.

The B cell EST array was custom spotted. Inserts from the selected bursal cDNA clones were PCR amplified using vector (pSport1) specific primers. The Arabidopsis negative control genes psbA, psbC and psbD were amplified from minipreps with vector-specific primers (M13 and M13R). These PCR products were purified by using an ArrayIt 96-well PCR purification kit (TeleChem International, Sunnyvale, CA, USA). The B cell EST array was then produced by nylon membrane-based cDNA spotting (Eurogentec, Seraing, Belgium).

For spotting on the larger BursalEST array plasmid DNA was grown up and PCR amplified using primers to the linker sequences at RZPD (Deutsches Ressourcenzentrum für Genomforschung). PCR products were purified and spotted onto a nylon filter macroarray in duplicate within a 4x4 subarray. The array contains 36864 (192x192) features including 14592 cDNA clones (in duplicate), spike control spots (not used in hybridizations) and empty spots. The array contains all sequences sequenced from the dkfz426 library to date and therefore contains some sequences, such as EF-1α, in high redundancy. This can aid in the statistical analysis. The array can be hybridized several times and quality of spots on the array can be assessed by hybridizing with linker oligos to determine DNA content of a spot.

4.1.2. Annotation pipeline for the Bursal EST array (I)

The larger BursalEST array contains 14592 clones from the dkfz426 Bursal EST cDNA library (I). Since the cDNA library from the DKFZ (Deutsches Krebsforschungszentrum) was not full length and poly-dT priming was used in their making, many of the sequences did not extend into the protein coding area. In the absence of the chicken genome sequence, these genes could not be annotated by referring to other, better annotated species, such as the human. Therefore, a sequence annotation pipeline was developed that included clustering of the array probe sequences together with the available chicken sequences including sequences from other chicken EST sequencing efforts (Abdrakhmanov et al. 2000; Caldwell et al. 2005; Hubbard et al. 2005).

Sequences on the array were clustered with the JESAM software (Parsons & Rodriguez-Tome 2000) together with the other sequences. Before BLAST searches were made the sequences were filtered and masked with Paracel filtering package (PFP) in order to remove spurious hits, especially due to matches to the chicken repeat element cr1 (Wicker et al. 2005). Repeat sequences for chicken were obtained from the repbase repeat database (Jurka et al. 2005). If BLAST searches of the sequence spotted on the array yielded a high confidence hit, then that hit was selected as the annotation. Otherwise, the best BLAST hits from each cluster were used to improve on the direct

sequence annotations. The bursalEST array was submitted to ArrayExpress under the accession A-MEXP-155 and in the Gene Expression Omnibus (GEO) under the accession GPL10224 (Alinikula et al. 2010).

4.2. Sample growth conditions protocols

4.2.1. Animals (II)

Chickens of the inbred HB.2-strain were maintained at the animal facilities of the Department of Medical Microbiology and Immunology, University of Turku, Finland. Embryos were incubated in a ventilated and humidified incubator at 37 ˚C.

4.2.2. Cells and cell lines (I, II)

Primary immune cell from various stages of development were collected two weeks after hatching (Figure 7). The bursa of Fabricius was dissected and single cell suspensions were prepared (II, Koskela et al. 2003). Splenic mononuclear cells were isolated with Ficoll-Paque gradient centrifugation (Pharmacia, Uppsala, Sweden) (Koskela et al. 1998). E14 splenocytes were isolated (Nieminen et al. 2000). Germinal centres were isolated from the spleen of chickens immunized with sheep red blood cells (SRBC) (Smithyman et al. 1979). The harderian gland was isolated, dissected and the cells were prepared for total RNA isolation using TRIZOL-reagent.

DT40 B cells were maintained in RPMI 1640 supplemented with 10% FCS, 1% chicken serum, 50 μM β-mercaptoethanol, 2 μM L-glutamine, penicillin and streptomycin (I, II, Koskela et al. 2003). The cells were cultured at +40 ˚C with 5% CO_2 (Buerstedde & Takeda 1991). Other cell lines (II) were cultured as described (Liippo et al. 1999).

4.3. Sample treatment protocols

4.3.1. FACS and MACS sorting and analysis

The pre-bursal and bursal primary cells were purified using a chicken B cell surface marker Bu-1 (Koskela et al. 2003; Houssaint et al. 1989; Nieminen et al. 2000). The B-cell marker Bu-1 is recognized by the mouse monoclonal antibody (mAb) L22 against the ChB6a alloantigen. MACS cell sorting columns were used to isolate B and T cells from bursa and the spleen of 2-week-old chickens, respectively. The germinal centres form a capsule which enables their removal from the surrounding tissue and the Harderian gland also consists of almost pure plasma cells (Mansikka et al. 1989). No further separation was used.

The mAbs TCR1 and TCR2 (Southern Biotechnology Associates, Birmingham, AL, USA) were used to purify γδ T-cell and αVβ1 T-cells, respectively. The MACS separations were carried out according to the manufacturer's instructions using goat anti-mouse IgG MicroBeads and VS+ separation columns (Miltenyi Biotec, Bergisch Gladbach, Germany). The percentage of viable cells after MACS sorting was over 95% according to Trypan blue staining (Fluka Chemika, Buchs, Switzerland). The purity was over 99% by FACScan flow cytometer (Becton Dickinson, Mountain View, CA, USA). The ChB6a-positive fraction of pre-bursal stem cells was isolated from the E14 spleen using FACStarplus and fluorescence activated cell sorting after staining with the L22 mAb (Becton Dickinson, Mountainview, CA, USA). Live cells were gated according to forward and side light scatter analyses. Percentages of positive cells in three separate experiments exceeded 99%.

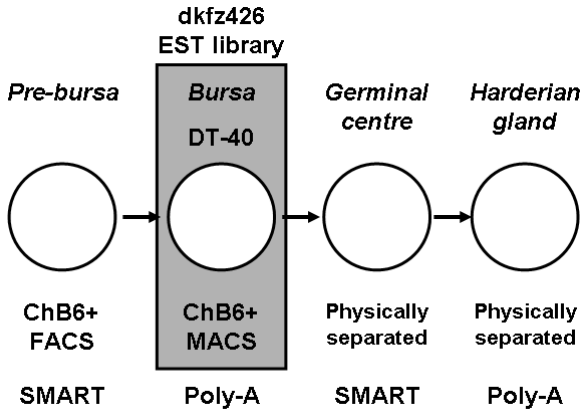

Figure 7. Avian B cell developmental hierarchy and extraction of samples for gene expression profiling (I). FACS indicates cell isolation using the FACStarplus instrument (Becton Dickinson, Mountainview, CA), MACS indicates isolation by magnetically activated cell sorting (Miltenyi Biotec, Bergisch Gladbach, Germany). Physical separation indicates no sorting. SMART indicates RNA isolation by SMART system (Clontech, Palo Alto, CA, USA), which includes amplification of the sample. Poly-A RNA was isolated using Trizol reagent (Life Technologies) and the Oligotex mRNA kit (Qiagen). In order to minimize effects of the isolation method on the results array analysis SMART RNA was only compared to SMART RNA and likewise with Poly-A RNA.

4.3.2. Pax5 gene inactivation in the DT40 B cell line (I, IV)

In order to inactivate the Pax5 gene in the DT40 bursal B cell line two Pax5 gene knockout constructs were made (I). In the gene-targeting construct *Pax5-neo* the 5' flanking arm was obtained by PCR from genomic DNA of DT40 cells by using the primers 2f and 3Lr, which were designed based on the coding sequence of chicken *Pax5* in the exons 2 and 3, respectively (Table 4.1). The 3Lr primer added *BglII* and *BclI* sites to the genomic PCR-product. An internal *BamHI* site within the *Pax5* intron 2 was used together with the *BglII* site (created by 3Lr) to clone the fragment into pUC18 vector. The 3' flanking arm was obtained by genomic PCR by using primers 3Rf and 4Rr designed for the exon 3 and 4 sequence of chicken *Pax5*, respectively. The 4Rr primer introduced a *BamHI* site, which was used for cloning together with the

internal *BclI* site within the *Pax5* intron 3. The created 3' flanking arm fragment was cloned the into the *BclI* site (created by 3Lr) of the 5' flanking arm that had been cloned into pUC18 vector. Finally, the neomycin resistance marker was cloned into the *BclI* site between the two flanking *Pax5* sequences.

Table 4.1. Primers and probes used in the making of the Pax5 knockout cells as well as Pax5, BCL6 and Blimp1 transfection constructs.

primer	***sequence (5' to 3')***
Pax5-2f	GTGAACCAGCTGGGGGGCGTTTTTGTGAAT
Pax5-neo-3Lr	TTTAGATCTGATCATCCAATCACTCCAGGCTAAATGCTCC
Pax5-neo-3Rf	ATTGCAGAGTA-CAAACGCCAAAATCCCACCA
Pax5-neo-4Rr	TTTGGATCCGGCTGCTGCACC-TTTGTCCGTATGAT
Pax5-bsr-b2Lf	CCGCGTCGACCACGGTACCGTCAGCTAAATACTCGGCA
Pax5-bsr-b3Lr	TGGTGTCGACCCTCCAATCACTCCAGGCTAAATGCTC
Pax5-bsr-b3Rf	AAAATTGCAGAGTACTAGTG-CCAAAATCCCACCA
Pax5-bsr-b4Rr	TTGGGCGGCCGCTGCACCTTTGTCCG-TATGAT
neo-f	GCGCA-TCGCCTTCTATCGCCTTCTTGACGAG
bsr-f	CGATTGAAGAACTCAT-TCCACTCAAATATACCC
Pax5-2p	GTCAGCCACGGCTGCGTCAGCAAAATACT
Px5-Hf	TATAAGCTTCGCAATGGATTTGGAGAAGATGTA
Px5-Br	TATAGATCTGCTTTGGGTCCGAGGTCAGTG
Bc6-Hf	AAAAAGCTTATGGCCTCACCGGCAGACAGCTGCA
Bc6-Nr	AAAGCTAGCTCAGCAAGCCTTGGGGAGCTCCGGA
B1-Nhf	AAAGCTAGCATGAAAATGGACATGGAGGATGCT
B1-Ncr	AAACCATGGTTAAGGGTCCATTGGTTCAACTGT

In the targeting construct *Pax5-bsr* the 5' flanking arm of the *Pax5-bsr* was obtained by genomic PCR by using primers b2Lf and b3Lr, which were designed based on the coding sequence of chicken *Pax5* exons 2 and 3, respectively. The b2Lf and the b3Lr primers both created the *SalI* sites, which were used to clone the genomic PCR-product into MCS I of the pLoxBsr vector (Arakawa et al., 2001). The 3' flanking arm of the *Pax5-bsr* construct was obtained by genomic PCR by using primers b3Rf and b4Rr designed for the exon 3 and 4 sequence of chicken *Pax5*, respectively. The b3Rf primer created a *SpeI* site and the b4Rr primer introduced a *NotI* site, which were used to clone the 3' flanking arm into MCS II of the pLoxBsr vector (Arakawa et al., 2001). The *Pax5-neo* and *Pax5-bsr* were linearized by *BamHI* and *Acc65I* digestion, respectively.

The constructs were introduced into DT40 cells by electroporation at 710 V, 25 μF. *Pax5-neo* was first transfected to wild type DT40 cells and heterozygous $Pax5^{+/-}$ mutant clones were then transfected by *Pax5-bsr*. Stable transfectant clones were selected in the presence of 2 mg/ml G418 (*Pax5-neo* transfectants) or 50 μg/ml blasticidin S (*Pax5-bsr* transfectants), respectively.

The clones that had incorporated the targeting construct into the correct genomic locus were identified among the transfection positive clones based on two genomic PCR

reactions with *Pax5* specific primer 2f used in combination with the selection cassette specific primer neo-f or bsr-f. The resulting genomic PCR-products were probed in Southern hybridization with the *Pax5* specific probe 2p.

4.4. RNA Extraction Protocols (I, IV)

In order to obtain RNA for array studies of the chicken B cells and T cells (Koskela et al. 2003) when starting material was scarce, the SMART RNA isolation and target amplification system (Clontech, Palo Alto, CA, USA) was used (Figure 7). The Poly-A RNA isolation method was used when cellular material was abundant, as indicated in Figure1. RNA was isolated with the Trizol reagent (Life Technologies) and mRNA was obtained with Oligotex mRNA kit (Qiagen) using magnetic beads bound with poly-dT probes.

The mRNA from wild-type DT40 and *Pax5*$^{-/-}$ cells (I, IV) was isolated using Trizol reagent (Life Technologies) and the Oligotex mRNA kit (Qiagen).

4.5. Labelling, hybridizations (I, IV)

Hybridization probes for the B cell EST array (Koskela et al. 2003) that were prepared from poly A+ mRNA isolated from MACS sorted bursal ChB6a+, splenic TCR1+ and TCR2+, Harderian gland and DT40 cells contained 0.5 to 1.0 μg of poly-A mRNA. Labelling was carried out using 1 μl of oligo (dT) 18 primer 0.5 mg/ml (Gibco), 2 μl of dNTP-mix (dATP, dGTP, dTTP) 10 mM, 1 μl of dCTP 0.1 mM, 4 μl of 0.1 M DTT (Gibco), 5 μl of 32P-dCTP 3000 Ci/mmole (NEN), 1 μl of RNase inhibitor (Promega) and 2 μl of Superscript II RT (Gibco) in total volume of 40 μl. The labelling reactions were incubated at 42 °C for 2 hours. Synthesis was stopped by adding 5 μl of 50 mM EDTA and 2 μl of 10N NaOH and further incubating 20 min at 65 °C. Subsequently, 4 μl of 5M acetic acid was added and probes were purified by spin columns (Clontech). Due to the small amount of mRNA in the sorted ChB6a+ cells from E14 spleens and isolated germinal centres, the cDNA-synthesis and its amplification were carried out using Atlas SMART cDNA-amplification kit according to the manufacturer's instructions (Clontech). To get better signal two hot radioactive nucleotides, 32P-dCTP and 32P-ATP, were used. After probe labelling and purification, radioactivity was measured with β-counter (1450 Microbeta PLUS liquid scintillation counter, Wallac, Finland). Non-stripped filters were pre-hybridized in 10 ml of ExpressHyb-solution (Clontech) containing 1 mg of ssDNA and 50 μl of SMART blocking solution for at least 2 hours. Labelled probes were added to the pre-hybridization solution and hybridized over night at 65 °C. Filters were washed four times with 2xSSC and 1% SDS for 30 min at a time and once with 0.1xSSC and 0.5% SDS for 30 min at 65 °C. Subsequently, filters were exposed to imaging plates for five to seven days and the resulting signals were captured on a phosphorimager (Fuji, Kanagawa, Japan).

Images of the B cell EST array (Koskela et al. 2003) were analyzed using P-Scan software (Carlisle et al. 2000) (abs.cit.nih.gov/pscan/) running on MatLab 5.1 (MathWorks Inc., Natick, MA, USA). Background was recorded above and below spots, individually for each spot. Too high spot background was normalized to the maximum ambient background. To avoid signal bleed into the background, the standard deviation of spot pair-wise ratios was recorded over all background spot pairs. When the ratio was larger than 2xSTD the lower value was taken as the background. Adjacent duplicate signal spot pairs were treated similarly. Background subtracted empty spots were used to calculate the standard deviation in signal measurements. Background subtracted signal intensities less than 2xSTD of the empty signal spots were floored to zero. If one spot in pair was zero, both were floored to zero.

Hybridization probes for the BursalEST array (I) were labelled using 0.5 μg mRNA. In order to carry out the labelling 0.5 μg dT(18)V (Eurogentech, Seraing, Belgium) was incubated 10 min at 70°C, 0.5 μl RNasin (40 U/μl, Promega, Madison, WI, USA), 5.0 μl reaction buffer, 2,5 μl 0,1M DTT, 0,5 μl 20 mM dGTP dATP dTTP mix (Promega), 5.0 μl α-33P-dCTP Ci/μl, Amersham Pharmacia Biotech Inc, Buckinghampshire, England, incubated 1 min at 37 °C, 1 μl Superscript II RT (Gibco BRL, Paisley, UK) and incubated 1.5h at 37 °C. The mRNA was hydrolysed by 1 μl 0.5 M EDTA, 1 μl 10% SDS, 3 μl 3N NaOH, incubated 30 min at 68 °C and 1 μl Tris-HCl (pH 8) and 3 μl 2N HCl were added. The labelled cDNA was purified using MobiSpin S-300 columns (MobiTec, Göttingen, Denmark) according to manufacturer's instructions.

In order to perform the hybridizations (I) the microarray membranes were pre-hybridized in Denhard's hybridization mix (12xSSC, 10xDenhard's solution (0.2% BSA, 0.2% Ficoll, 0.2% polyvinylpyrolodone), 0.5% SDS, 100micrograms/ml ssDNA) and 7.5 μl d(A)40 (Amersham) for 2 h at 65 °C. The labelled cDNA was added to the hybridization mix and hybridized to the membranes over night at 65 °C. The membranes were washed at 65 °C 20 min in 1XSSC, 0.1% SDS, 2x10 min 0.3xSSC, 0.1xSDS and 0.1xSSC and 0.1%SDS. The radioactive signals were read using phosphoimager plates and a phosphoimage-reader Fluorescent Image Analyzer FLA-3000 (Fuji, Kanagawa, Japan). The scanning software supplied by the manufacturer with default settings was used for scanning producing a BAS image (*.img) file. Scanning resolution of 50 μm was used (the smallest available). The quality of the spots on the array was assessed by hybridizing with linker oligos to determine DNA content of a spot.

Image analysis of the BursalEST array was carried out using the AIDA image analyzer v.3.27 software (I). Empty spots in the array were employed to calculate the background. Occasionally these background adjustments are skewed by signal bleed from neighbouring high signal intensity spots. This creates an uneven background signal. A custom visual basic protocol, employing a sliding 4x4 grid, was used to smooth the background signal.

RNA samples from the wild-type and Pax5-/- cells were hybridized to the Chicken GeneChip array (Affymetrix, Santa Clara, CA, USA) (IV). Sample processing and

labeling were performed according to the protocol provided by Affymetrix. Chips were scanned using the GeneChip Scanner 3000 (Affymetrix).

4.6. Array data analysis protocols

4.6.1. Normalization and statistical analysis of B cell EST arrays

Normalization of the B cell EST arrays, which is able to assay 230 B-cell specific genes (Koskela et al. 2003), was done with TIGR ArrayViewer (www.tigr.org) using the ratio statistics method as described (Chen et al. 1997). To determine whether a gene was differentially expressed, 2-3 replicate experiments were made and all spots were individually analyzed. If 5 of 6 or 4 of 4 spots were differentially expressed by more than two-fold, then a randomized set of self vs. self comparisons indicated that gene to be more than 98% likely to be differentially expressed i.e. the permutation derived p-value was less than 0.025. The log-ratios of median centred spot intensities were used for clustering. Clustering was done with J-Express (Dysvik & Jonassen 2001) (MolMine AG, Norway).

4.6.2. Analysis of Bursal EST arrays (I)

Pre-normalization filtering was employed to correct for uneven spotting of the arrays and to focus analysis on those spots which had a signal above the background level in either the wild type or knockout cells. Criteria for including the spots in the analysis were: 1.) Oligonucleotide hybridizations had to show DNA in the spots that were included in the analysis. If oligo hybridizations did not detect DNA (defined as 2xST_DEV of global spot background) the spot was excluded from analysis. If either of the two technical replicate spots was excluded in this manner both were excluded. 2.) If in the Pax5 versus WT hybridizations too few spots were present (less than 1/3) or spots were present in higher quality hybridizations but absent in the low quality hybridizations in various combinations the spots were also excluded from the analysis. After the filtering steps, 6735 out of the 14492 spot-pairs remained, where there was DNA in the spots and the gene was likely to be expressed in either the knockout or the wild-type experiments.

A correction for signal bleeding artifacts was based on the idea that acquiring signal from a neighboring high-intensity spot would lead to a significantly higher signal than the technical replicate of that clone. Oligonucleotide hybridizations were employed to define what would be a significantly higher ratio. The following formulas were used: abs (Rep1/Rep2 - 1) < 3 x standard deviation (oligo-pair-ratios). The signal value of the bleed-positive spots was replaced with an "NA" and k-means method employing the ten most similar genes was used to impute a new value for the measurement. The Significance analysis of microarrays (SAM) package was used for this task (Tusher et al. 2001).

The maanova R-based utility was used for normalization (Kerr et al. 2000). A combined spatial 2D loess and intensity dependent loess approach was employed. Variance stabilizing normalization (VSN) (Huber et al., 2002) from the R/Bioconductor package vsn (Gentleman et al., 2004) was introduced in order to improve performance of the SAM package in distinguishing differential expression (Tusher et al. 2001). A reasonable false discovery rate was chosen (about 5%) and results were tabulated for further annotation and verification with quantitative PCR. The descriptions of the experiments and the data are available at the ArrayExpress (EBI, Hinxton, UK) (Parkinson et al. 2009) under accession E-MEXP-270.

4.6.3. Normalization of the Chicken Affymetrix Genechip array (IV)

The raw *.cel files were processed separately using the R language (R Development Core Team 2009) and the RMA method implemented in the Bioconductor (Gentleman et al. 2004; R Development Core Team 2009) package affy (Irizarry et al. 2003). The microarray data were deposited in the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo) database. For meta-analysis Log2-ratios of the Pax5-/- and wild-type expression results were calculated for the three biological replicates and used in subsequent comparisons.

4.7. Data verification protocols

4.7.1. Semi-quantitative RT-PCR and RT-PCR (I, II)

For semi-quantitative reverse transcriptase PCR (RT-PCR) analyses, RNA isolation was performed with TRIZOL and cDNA synthesis was performed as described earlier (Koskela et al. 1998). Primers are listed in Table 4.2. Following an initial 2 min denaturation step at 95 °C, the PCR conditions were 95 °C for 30 s, 54– 60 °C (depending on the primer pairs) for 30 s, 72 °C for 1 min for 22, 24, 26, 28, 30, 32, 34 or 36 cycles. Products were loaded onto a 1.5% SeaKem agarose gel (FMC Bioproducts, Rockland, Me.).

Table 4.2. Primers and probes used for semi-quantitative RT-PCR in Koskela et al. 2003.

Gene	*forward primer/probe (5' to 3')*	*reverse primer (5' to 3')*
IgL chain	TCAGGTTCCCTGGTGCAGGCA	TGCTGTGGTCTCGCCAGA GC
BCAP	TATGGGCTGAAAAACCTAACTGCT	CATCCGTGCTACCTCCTTCTAAAA
Cyclin B2	TGTGTTGCAGTTATGGACCGCTTC	ACTGCTGCTTTG TACCCCACTTATCA
bcl-x	CCG ACC ATC TAG ATC CCT GGA	GTG GAT GTG TGA AGG CGC AGC
SWAP-70	GGAGCTAGAAAACCAGAGAATGATA	GGG CCC CAGTTTGTGATGAG
CIIDBP	ACTGGGGCACGGTGAGGTTTG	GGAAGAGGCGCAGCACGGTAGTAT
CXCR4	GGA CGG CCC GGA CCT ACT CG	GGCAGCCAGACACCCACATACACA
β-actin	GTGCTGTGTTCCCATCTATCGT	TGG ACAATGGAGGGTCCGGATT

Pax5 expression of wild-type DT40 , Pax5+/-, and Pax5-/- clones was analyzed by RT-PCR (I). The cDNA from $1x10^5$ cell equivalents, as described previously (Nera et al. 2006b), was amplified with primers Pax5-f and Pax5-r (Table 4.3). A PCR reaction with the chicken b-actin-specific primers b1-f and b1-r was used as a positive control. Southern hybridization was performed with the Pax5-specific probe Pax5-p (Table 4.3).

Table 4.3. Primers and probes used for RT-PCR in Article I

Gene	***forward primer/probe (5' to 3')***
Pax5-f	GTCAGCCACGGCTGCGTCAGCAAAATAC
Pax5-r	GGCTGCTGCACCTTTGTCCGTATGAT
Pax5-p	ATTAAGCCTGGAGTGATTGGAGGATCAA
□-actin-f	GTGCTGTGTTCCCATCTATCGT
□-actin-r	TGGACAATGGAGGGTCCGGATT

Total RNA was isolated for RT-PCR analysis (II) using the Ultra- SpecTM-II RNA Kit (Biotecs Laboratories, Houston, TX, USA). Oligo-p(dT)-primed cDNA were synthesized using avian myeloblastosis virus (AMV) reverse transcriptase (1st Strand cDNA Synthesis Kit for RT-PCR, Boehringer Mannheim, Mannheim, Germany). PCR primers were Helios-exon2f1 and Helios-exon6r (Table 4.4). β-actin was amplified from the same pool of cDNA (Liippo et al. 1997). In order to amplify exon 4b containing isoforms, Helios-exon2f primer was used with Helios-exon 4b-r. Isoforms lacking the exon 6 were amplified with Helios-exon2f2 and Helios-exon7r. Products were separated by gel electrophoresis, transferred to Hybond™-N nylon membrane (Amersham Int.,Amersham, UK) and hybridized with γ32P-ATP-labelled internal Helios-exon4a to show all isoforms and with Helios-exon4b to show only those isoforms containing the novel exon. Hybridization was carried out at 58 °C (Liippo et al 1999).

Table 4.4. Primers and probes used in Article II.

primer	***sequence (5' to 3')***
Helios-exon2f1	CCTCACTGAGAATAACGAGAT
Helios-exon6r	CTTCTCTATAACAGCAGGTCTCT
Helios-exon 4b-r	TCCTGCAGTGCCATCTCTTATA
Helios-exon2f2	GACATGGCGGATAACAGAAAGAT
Helios-exon7r	ATAGACCTGAGCATAAGGTGA
Helios-exon4a	CACTGCAACCAATGTGGAGCTTCTTTTA
Helios-exon4b	TGAACTACAGCAGTGAGCTCCTGTATAT
Helios-mmF	AAGTCATGACAAGCACAAATT
Helios-mmR	CAAGGTAGGTGATTGCATTGTT
Helios-ch2R	GTCTTCCTGTCAGATTCCTCATCACTTT
pCDM8-F	GCAGAGCTCTCTGGCTAACTA
Helios-ch6F	GCTGACAAGCAACTTGGGAAAGCGTAAAA

4.7.2. Long-range PCR to investigate structure of Ikaros family genes (II)

In order to investigate the gene structure of the chicken Helios gene the Expand™ Long Template PCR System (Boehringer Mannheim) was used to amplify intron sequences from chicken genomic DNA prepared according to the manufacturer's instructions. Primers (28- to 31-mer) were designed to anneal at the exon/intron boundaries. The PCR products were separated out on a 0.5% agarose gel. The genomic structure of chicken Ikaros and Aiolos was investigated previously (Liippo et al. 1997 and 1999).

4.7.3. Light Cycler ™ quantitative RT-PCR (I)

The quantitative real-time PCR analysis was made using LightCycler equipment (Roche) and the SYBR Green detection method. The LightCycler analysis was made using LightCycler FastStart DNA Master SYBR Green I kit (Roche) according to the manufacturers' instructions. 2 µl of serial template dilutions were used. The Mg^{2+} concentration and PCR conditions were optimised for each primer pair (Table 4.5). Melting curve detection was run after each analysis. GAPDH, β-actin and elongation factor 1 (EF-1) were used to normalize for the cDNA concentration between samples.

Table 4.5. Quantitative RT-PCR primers for Article I and Koskela et al. 2003.

GENE	***forward-primer (5' to 3')***	***reverse-primer (5' to 3')***
CD79b	GCGTCCCCATGCTCCTCTTCCT	GCAGCACCCCTCACTCCTCTCCT
BLNK	CTTCCCCTCACAGCAGCATTTCAT	AACGCTCTTCACCACATTTTTCTC
Lyn	GATGTGATGGTTGCTCTTC	GGTCGGCCTTTCTTCTG
Btk	GAGG-CAGCAAGAAGGGCT	TGCACCAGGTCGCTGTTGTAT
VpreB3	GAGGCCTTGTCCCTTCTTTC	GCTACCA-AAGACCTCGTCCA
IgL λ	TCAGGTTCCCTGGTGCAGGCA	TGCTGTGGTCTCGCCAGAGC
HSP70	CAATGGCAAAGAGCTGAACA	GGTGGGAATG-GTGGTGTTAC
ITM2A	CTGACATTCGGGAGGATGAT	CGCCAGTTTTGCAAAGAGAT
HMG-17	AACCCTGCAGAAAATGGAGA	CAAAGCGGTCTGTGTGCTAA
Caspase 3	GCAGACAGTGGACCAGATGA	CTGCCACTCTGCGATTTACA
XBP-1	GTGCGAGTCTACGGATGTGA	AAGCCGAACAGGAGATCAGA
Blimp-1	ACACAGCGGAGAGAGACCAT	GCACAGCTTGC-ACTGGTAAG
BCL6	GAGAAGCCATACCCCTGTGA	TGCACCTTGGTGTTTGTGAT
IgH µS	GGAGAACCCCGAAAATGAGT	GCCAACACCAAGGAGACATT
IgH µM	GGAGAACCCCGAAAATGAGT	GTTGGATGTCGTCGTCCTCT
AID	GTTTCTGTGCACCAGAGGGCTGAACAGTCA	CTCCTTTCTTGGCTGGGTGAGAGGTCCATA
Aiolos	GGGATGCGCTAACGGGTCACTT	GTCGCCTTC-CTTCTCATACACG
EBF	GTGCAGCCGCTGTTGTGACAA	GGGACCATAGTCAATGGTGGG
GADPH	GAGGTGCTGCCCAGAACATCATC	CCCGCATCAAAGGT-GGAGGAAT
β-actin	AAGCCAACAGAGAGAAGATGACACA	TACAGATCCTTACGGATATCCACAT
EF-1	GGTTATGCCCCTGTGCTGGATT	CTTCTTGTCGACGGCCTTGATGA
ChB6	GCA GTA GAG CCT GGG GAA ATG	CCA AGG TCC TGA AGC CAC ATG
BAFF	CAT TGT CCC TTG GCT TCT GAG	TTC CTG TTT TGG GTT TGC TTA TTA
Pax5	GAA CGA GTG TGC GAT AAC GAC A	TCG CGA CCT GTT ACG ATA GGA

4.7.4. Western Blotting (I)

Wild-type, *Pax5*$^{-/-}$ and *Pax5*$^{-/-}$/Pax5 cells (4 x 10^7 of each) were washed twice with PBS at room temperature. Each cell sample was suspended to 1 ml of the methionine and cysteine free DMEM medium (Gibco) supplemented with 10% dialyzed FCS and was incubated 30 min at +40 °C. Next, 200 μCi/ml of the Redivue™ PRO-MIX™ L-[^{35}S] *in vitro* cell labelling mix (Amersham Biosciences) was added and the cells were incubated for 15 min at +40 °C. Following this, 4 ml of normal DMEM medium (Gibco) containing 10% FCS and an excess (5 mM) of L-cysteine (Sigma) and L-methionine (Sigma) was added, and each sample was divided into five independent 1 ml sample cultures (each containing 8 x 10^6 cells/ml), which were incubated for the indicated chase times (0, 30 min, 1 h, 2 h and 4 h). After incubation, the carefully cleared supernatants from the samples were subjected to the IgM immunoprecipitation.

4.8. Other protocols

4.8.1. BCR signaling via Ca2+ flux analysis (I)

Cells (10^6) were suspended in buffered solution containing 20 mM Hepes, 5 mM glucose, 1 mM $CaCl_2$, 0.25 g/l BSA and 0.25 mM sulfinpyrazone (Sigma) in PBS, pH 7.4, and were loaded with 3 μM Fluo-3 AM (Molecular Probes) for 45 min at room temperature. Following the loading period cells were washed three times and incubated an additional 30 min to ensure the complete cleavage of acetoxymethyl group from Fluo-3. Cells were washed twice and continuous monitoring of fluorescence from cell suspension (1 x 10^6/ml) was performed at +37 °C using a FacsCalibur flow cytometer at the excitation wavelength of 488 nm and an emission wavelength of 530 nm. Cells were stimulated with 4 μg/ml of M4 mAb. The average signal curve was measured by calculating the average of the fluorescence value of the events at every time point.

4.8.2. Pulse-chase metabolic labelling and IgM secretion analysis (I)

For immunoprecipitation, 50 μg of anti-IgM mAb (M1) were conjugated to 500 μl of protein A/G-PLUS-agarose reagent (Santa Cruz Biotechnology) for 12 h at +4 °C in 1 ml of lysis buffer (1 x PBS, 1% Nonidet P-40, 0.5% sodiumdeoxycholate, 1% SDS, 1mM EDTA, 2 mM phenylmethylsulfonylfluoride, 1 mM Na_3VO_4 and 1 x protease inhibitor 'cocktail' (Roche)) and then saturated for 2 h in PBS containing 5% BSA. After two washes in lysis buffer, the conjugated M1 antibodies were suspended to 500 μl of the lysis buffer. 20μl of this conjugated M1 suspension was added to each cleared supernatant sample in immunoprecipitation and incubated for 12 h at +4 °C. The immunoprecipitates were washed four times with the lysis buffer; the samples were denatured at +75 °C for 10 min in SDS sample buffer and separated by 4-12% SDS-PAGE. The radioactive gels were fixed in 15% methanol, 7.5% acetic acid and treated with Enlightning™ autoradiography enhancer (PerkinElmer) before drying and exposure on autoradiographic film.

4.8.3. Pax5, BCL6 and Blimp1 over-expression constructs (I)

In order to make constructs for over-expressing B and plasma cell regulators in DT40 cells, coding sequences of various genes were amplified from DT40 cDNA. The primers Px5-Hf and Px5-Br; Bc6-Hf and Bc6-Nr; and B1-Nhf and B1-Ncr were used for Pax5, BCL6, and Blimp-1, respectively. PCR products were cloned into the pExpress vector (Arakawa et al., 2001) and the insert was sequenced. The expression cassettes containing the cloned PCR products between the chicken β-actin promoter and the SV40 poly-A sequence were excised from pExpress as SpeI cassettes, which were subsequently cloned into pLoxPuro (Pax5 and BCL6) or pLoxHisD (Blimp-1) vectors (Arakawa et al., 2001). The vectors were linearized and transfected to the Pax5-/- (Pax5 and BCL6), wild-type, or Pax5-/-/Pax5 (Blimp-1) cells at 710 V, 25 mF. Stable transfectants were selected in the presence of 0.5 mg/ml puromycin (Pax5 and BCL6) or 1 mg/ml histidinol (Blimp-1), and the expression of transfected gene was verified by immunoblots (Pax5) or RT-PCR (BCL6 and Blimp-1).

4.8.4. Cloning of the Helios gene (II)

The primary oligonucleotide primers were designed from the mouse *Helios* cDNA sequence: 5'- Helios-mmF and Helios-mmR. PCR was done with DyNAzyme™ (Finnzymes OY, Espoo Finland). Amplifications on thymic cDNA yielded a fragment that was cloned using the Original TA cloning kit (Invitrogen, Carlsbad, CA), sequenced using the ABI PRISM™ Dye Terminator Cycle Sequencing Ready Reaction Kit with AmpliTaq® DNA Polymerase, FS (Perkin Elmer) and ABI 373 DNA Sequencer (Applied Biosystems Inc., Foster City, CA).

The 5'- and 3'- ends were obtained using vector- and *Helios* specific primers in a PCR reaction on cDNA (cloned into pCDM8 vector) derived from an inbred RPL-Line 0 chicken thymus (Liippo et al. 1997). A linear amplification at a higher temperature with one 28- to 31-mer gene specific primer was followed by an exponential amplification at the specific annealing temperature of the vector-specific primer (21-mer). This resulted in lower background and longer specific products than without the linear amplification. The 5'-reaction was done with Helios-ch2R and pCDM8-F; the 3'-reaction with Helios-ch6F using the vector primer as the reverse primer (Table 4). Thermal cycling conditions were 95 °C 2 min; 30 cycles of 95 °C 30 s, 65 °C 30 s and 72 °C 2.5 min followed by 25 cycles of 95 °C 30 s, 55 °C 30 s and 72 °C 1.5 min and finally at 72 °C for 30 min. 0.6 U of enzyme was used in 50 µl reaction volume, 75 µM dNTP and 15 pmol of each primer. Specific PCR products 2.1 kb (5'-end) and 1.7 kb (3'-end) in length were generated. These were TA-cloned and sequenced. The sequence was verified by sequencing multiple independent PCR products spanning the length of the gene.

4.9. Phylogenetic analysis protocols (III)

Various programs were used to infer a dendrogram based on maximum likelihood distances. Sequences for several species were downloaded. These included lamprey

(IKLF1; AAL67302 and IKLF2; AAL62094 and AAL67304), hagfish (Ikaros-like; AAP84653), mouse (Eos; NP_035902, Helios; NP_035900, Ikaros; NP_033604 and Aiolos; XP_283022), human (Eos; NP_071910, Helios; NP_057344 and Ikaros; NP_006051), chicken (Ikaros; O42410 and Aiolos; CAB56282), skate (Helios; AAF87270, Ikaros; AAF87271 and Aiolos; AAF87273), the axolotl frog (Ikaros; AAF01038), newt (Ikaros; CAC84566), trout (Ikaros; AAB53434) and zebra fish (Ikaros; NP_571061, Eos; ENSDARG00000003885). Chicken Helios (CAC59948) was truncated to remove the exon 4b which appears to be chicken specific. Furthermore the ensembl database was searched for novel fugu fish ikaros family members. Four were identified: SINFRUP00000146537; closest to Ikaros, SINFRUP00000147157; closest to Helios, SINFRUP00000162634; closest to Eos and SINFRUP00000158625; also closest to Eos. There may be more fugu Ikaros family members yet to be identified.

The program T-Coffee (v. 1.37) (Notredame et al. 2000) was used for sequence alignment. Slow and accurate settings were employed throughout. BioEdit (Tom Hall, North Carolina State Univ., USA) was used to truncate sequences removing the first zinc finger. Also columns with gaps were removed because the gap has no definition in the amino acid substitution tables. PHYLIP (version 3.6a3) program seqboot was used to produce a bootstrap set with 1000 replicates (Felsenstein 2004). The program Treepuzzle (v. 5.1) and the puzzleboot script (v. 1.03) were then used to infer maximum likelihood distances with gamma correction (6-8 variant and one invariant rate categories) (Schmidt et al. 2002). The VT substitution matrix was employed (Muller et al. 2000) and only sequences that passed the chi-square test for frequency distribution assumed in the maximum likelihood model were included in the study. Phylip was used to infer a phylogenetic tree from the distance matrices. A dendrogram based on these distances was constructed using the Fitch-Margoliash method in the Fitch-program. The global rearrangements option and 10-fold jumbling were used to maximize the changes of obtaining the optimum tree. The TreeView (version 1.6.6) was used to display the trees (Page 1996). For the second tree Pegasus sequences were obtained for use as outgroups from human (NP_071911), mouse (NP_780324) and rat (XP_219325) (Perdomo et al. 2000).

4.10. Cross-species meta-analysis protocols (IV)

4.10.1. Data Collection and annotation

Three datasets from murine Pax5 knock-out experiments were obtained from the GEO. Twelve samples of wild-type versus Pax5-/- global gene expression profiling experiments were downloaded (Cobaleda et al. 2007a; Schebesta et al. 2007). The samples were from the superseries GSE8461 composed of series GSE8457 and GSE8458 containing *in vitro* cultured and *ex vivo* sorted murine pro-B cells, respectively. Platforms GPL5518 and GPL5519 from the IMP Vienna microarray facility were used in both series. The third murine dataset came from the profiling of

Pax5 +/+, Pax5 -/-, and Rag1-/- pro B cell lines (Pridans et al. 2008) and was part of the series GSE9345. The series GSE9345 utilized the NIA15k gene expression platform with the GEO accession GPL5990. Since biological differences between the two subseries of the GSE8461 superseries appeared minor, the data were grouped and analyzed according to the platform. The data were obtained in normalized format, typically after loess normalization and log2 transformation. Missing values in the data were marked with NA.

The Affymetrix probeSet_IDs were mapped to the orthologous mouse and human Affymetrix probeSet_IDs using the Chicken.na30.ortholog file obtained from the NetAffix™ Analysis Center (Affymetrix, http://www.affymetrix.com). Gene symbols corresponding to the human and mouse probeSets were obtained from the Stanford SOURCE database (http://source.stanford.edu) (Diehn et al. 2003). In order derive results comparable to the mouse data, the mouse orthologs were first inspected, then the human and finally chicken-derived gene symbols were used. The mouse platforms were cDNA glass slide microarrays that used dbEST sequence identifiers to annotate probes. The SOURCE database was used to obtain gene symbol references for these probes as well.

4.10.2. Creation of gene expression compendia

Platform, organism and over-all gene expression compendia were generated. The data were first treated at the platform level. Probes containing over 50% missing values were removed from the analysis. The log2 ratios of the normalized expression values were then summarized at the gene level using Huber's M-estimator of location (Huber 1981) from the MASS R package (Venables & Ripley 2002). The mouse compendium was generated in the same way, except only 33% of missing values were tolerated. Genes selected for the over-all compendium had to contain more than half of the values present in more than half of the platforms. So as to enable the cross-species analysis, the over-all compendium contained only genes that were present on the genome-wide chicken affymetrix array. The statistical analyses were carried out using tables with NA values replaced by row-wise means. Visualization and inspection of the results was done using tables without any imputed values.

4.10.3. Differential expression analysis between mouse and chicken

In order to obtain a clearer picture of the biological differences and to remove genes that clearly behaved in a different way between the organisms, a differential expression analysis between mouse and chicken was carried out. As the data had not been comprehensively normalized so as to allow the use of location tests, such as the t-test, rank-based methods were employed. The RankProd R-package (Hong et al. 2006) was used. These differentially expressed genes reflect the differentiation status *vis a vis* mouse samples representing early B-cell differentiation and the DT40 Pax5 -/- cells having undergone plasma cell transition.

4.10.4. Rank-based cross-species meta-analysis at the gene level

After removal of the differentially expressed genes between the chicken and the mouse, a rank-based meta-analysis of the three mouse platforms and the chicken platform was performed. The analysis was carried out, using the RankProd package RPadvance function, in a balanced fashion so that rank-products were first calculated for each species separately and then combined across species. The genes that were similarly regulated in chicken and mouse after Pax5 removal were obtained. The lists of co-ordinately regulated genes obtained without first removing the differentially expressed genes between the organisms were compared to the results of this analysis and found to be largely similar (data not shown).

For visualization purposes the genes expressed in a similar fashion across species, and those differentially expressed between species, were combined and clustered together. An in-house heatmap.n R-function was used which splits the genes into a pre-defined number of groups showing similar expression using the partitioning around medoids (PAM) method (Kaufman & Rousseeuw, 1990) from the cluster R-package, and then performs hierarchical clustering within the groups.

4.10.5. Gene Set Enrichment Analysis

Gene set enrichment analysis (Subramanian et al., 2005) was performed on all of the four individual data sets as well as for the mouse and chicken compendia. The Gene Set Analysis (GSA) R package (Efron & Tibshirani, 2007) was used for defining the enriched gene sets in the data. Gene sets were obtained from the Molecular Signatures Database (MSigDB, Broad Institute of MIT and Harvard, Cambridge, MA) (Subramanian et al., 2005). The MSigDB curated gene sets (C2, 1892 sets) and motif gene sets (C3, 837 sets) including the 3'-UTR miRNA binding motifs (222 sets) (Xie et al., 2005) were downloaded. The ‘‘Maxmean’’ statistic was used to calculate enrichment scores, and permutation based p-values were derived from 100 bootstrap replicates. A false discovery rate (FDR) correction was also applied.

Differential expression at the gene level was also measured individually on each of the data sets used in the GSEA. Statistical analysis of differential gene expression was performed with R/Bioconductor (Gentleman et al., 2004) using the limma package (Smyth, 2005). Gene expression was compared to wild-type negative control in a pairwise fashion using the empirical Bayes statistics implemented by eBayes function (Smyth, 2004). The threshold for differential expression was set at $q<0.05$ after the Benjamini-Hochberg multiple testing correction. In order to display the differentially expressed genes, hierarchical clustering and PAM partitioning was performed with R (Kaufman & Rousseeuw, 1990). Results from these analyses were compared with the cross-species rank-based meta-analysis.

4.10.6. Cross-species pathway meta-analysis

The scores and p-values from the GSEA analysis above were saved and the scores from the four analyses were formulated into an enrichment scores matrix while

removing the NA rows, as above. A rank-based meta-analysis of the scores across the datasets was performed using the RankProd R-package (Hong et al. 2006). In order to weed out imbalanced sets where there was high enrichment in only one organism, a further requirement of a minimum score of at least +/- 0.1 was imposed. The results were visualized using the heatmap.n function (see above). Results from the cross-species analysis were compared to the within-species GSEA analysis (Tables 5.6-5.7), see below.

GSEA results from the organism-specific compendia were carried out in order to ascertain how much meta-analysis adds to the picture (IV Supplementary Figure 1). In addition, the results obtained from the organism-specific analyses were compared to the pathway meta-analysis results (Tables 5.6-5.7).

In order to connect the gene sets to each other, overlaps between gene sets using Fisher's test were calculated using the tools at the MsigDB website (www.broadinstitute.org/gsea/msigdb). Results of the overlap analysis are displayed in Supplementary Table 2. A significance threshold of p.value < 0.05 was used.

5. RESULTS

5.1. Development of the avian array platforms (I)

Since at the time ready-made arrays to perform gene-expression analysis on chicken did not exist, these tools had to be developed. Two array platforms were developed (Figure 8). Both are based on the Bursal EST cDNA library and database (Abdrakhmanov et al. 2000). The smaller platform (Koskela et al. 2003), referred to as the B cell EST array, contained a set of genes that was enriched for genes that are preferentially expressed by the B cells. The larger array (I) contains almost the entire bursalEST database on a filter array.

A pipeline was developed for annotating the BursalEST array (Figure 8). It employs methods similar to those used by the NCBI Unigene database (Pontius et al. 2003). Instead of using BLAST (Altschul et al. 1997), a method referred to as JESAM (Parsons and Rodriguez-Tome, 2000) was used to carry out EST clustering.

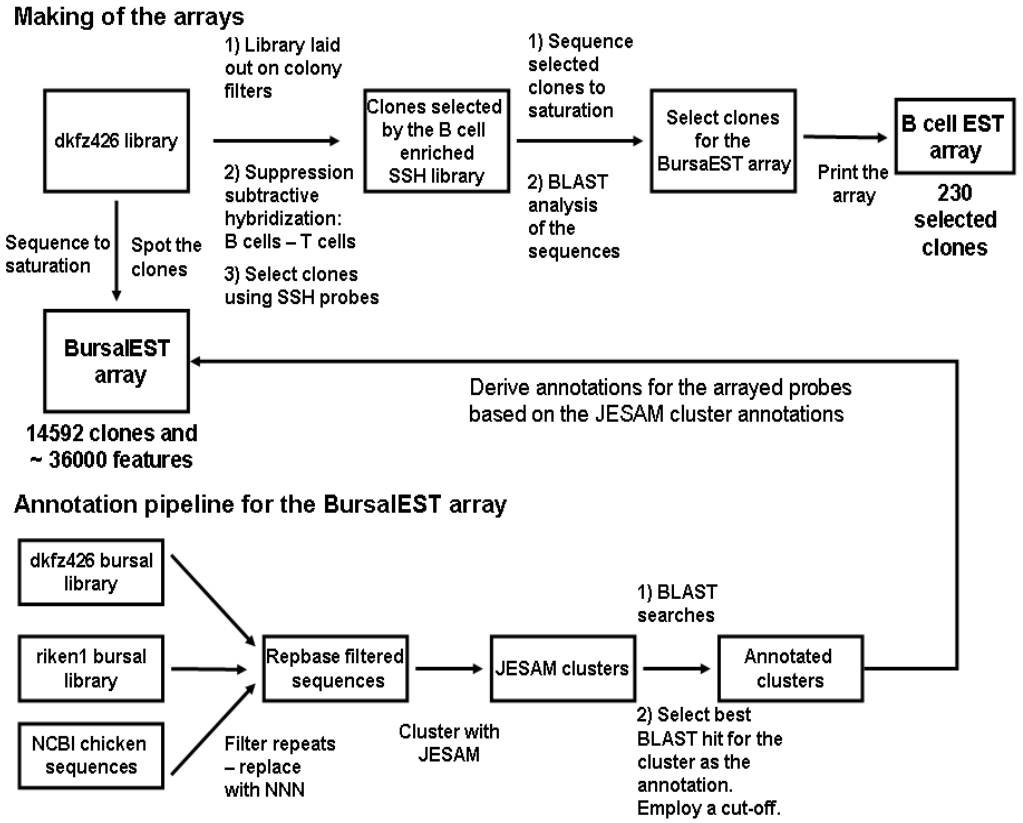

Figure 8. Making and annotation of the B cell EST and BursalEST array platforms (I).

5.2. Development of the avian array analysis methods (I, IV)

Several workflows were developed to carry out the data analysis (I, IV, Koskela et al. 2003, Koskela et al. 2004, Suonpää et al. 2005). Data extraction and normalization methods for the custom arrays also needed to be developed (I, Koskela et al. 2003, Alinikula et al. 2010).

Freely available software for reading in numerical values from the smaller B-cell EST array phosphorimager images was obtained (Koskela et al. 2003; Carlisle et al. 2000). Since the large BursalEST filter array (I, Alinikula et al. 2010) had surface irregularities that differed from array to array, various surface normalization methods were tested that could normalize between arrays. The maanova R-based utility was adopted for this purpose (Wu et al. 2003), since it is able to carry out surface loess normalizations and as the R/Bioconductor package communicated well with other methods. A combined spatial 2D loess and intensity dependent loess approach was selected as the method that produced the most reproducible results.

For the cross-species meta-analysis (IV) the chicken Affymetrix platform needed to be directly compared to mouse cDNA-clone based expression platforms. In order to map probes across chicken and mouse, the chicken Affymetrix probeset IDs were mapped to the orthologous mouse and human Affymetrix probeset IDs using mappings obtained from the NetAffix™ Analysis Center (Affymetrix, http://www.affymetrix.com). The Affymetrix ids were then mapped to mouse gene symbols.

In order to perform the cross-species gene expression meta-analysis, the gene expression compendia (IV) needed to have only one expression value per gene symbol. Two methods for summarizing or combining log-ratios were tested, namely Huber's M-estimator of location (Huber 1981) and the Tukey bi-weight method from the affy R/Bioconductor package (Irizarry et al. 2003). Huber's method, which was adopted, was found to increase slightly the number of top-scoring genes when compared to the non-summarized results.

In order to carry out the cross-species meta-analysis (IV), two data-analysis work-flows were built using custom R code and packages from the R/Bioconductor environment (Gentleman et al. 2004). The first, termed the GSA_eBayes workflow, combines gene set enrichment analysis (GSEA) using the GSA R package (Efron & Tibshirani 2007) with two-class differential gene expression analysis based on the empirical Bayes (eBayes) methodology from the limma R/Bioconductor package (Smyth 2004). The topmost differentially expressed gene sets are visualized with heatmaps. Differentially expressed gene sets contain some genes which are not differentially expressed. But only the differentially expressed genes were included in the gene-level heatmap visualizations of the topmost gene sets.

The second workflow, RankProduct-based data analysis, takes a matrix of values and performs rank-based differential expression analysis using the R/Bioconductor package RankProd (IV, Hong et al. 2006). It further performs visualization and clustering of the output. These two workflows were combined with adjoining R-routines and run in

series in order to carry out rank-based meta-analysis of the GSEA pathway enrichment analyses from both mouse and chicken species and all the platforms.

5.3. Bursal B cells have a similar expression profile to the DT40 cell line (III)

The B-cell EST array (Koskela et al. 2003) was used to profile the expression of a selected set of B cell enriched genes in avian B cells, T cells and plasma cells from the Harderian gland. Signatures associated with each of the cell types were established and compared to studies done on human and mouse. This helped to characterize the avian-specific B cell organ bursa of Fabricius at the level of the gene expression. Furthermore, the pre-Bursal cells were shown to express genes related to B-cell receptor signalling, even though they do not express the B-cell receptor on the cell surface (Koskela et al. 2003, Figure 4a). Selected differentially expressed genes were verified using semi-quantitative and LightCycler™ quantitative RT-PCR (Table 5.1). The Bursal B cells and the DT40 cell line also express the BCL6 gene, which is associated with germinal centre (GC) B cells in the mammalian B cells. The B-cell associated survival factor BAFF was also found to be expressed by bursal B cells.

Table 5.1. Genes that are differentially expressed between bursal B cells and TCR2+ T cells. sq-RT-PCR refers to semi-quantitative RT-PCR and q-RT-PCR_FC refers to fold-change obtained from quantitative RT-PCR between Bu-1+ bursal B cells and TCR2+ splenic T cells.

Gene Name	Symbol	Clone ID	LFC	# Changing	p-value	sq-RT-PCR	q-RT-PCR_FC
IgL chain	IGL@	dkfz426_5g3r1	7.91	6	1.328E-04	x	
BCAP	PIK3AP1	dkfz426_49p15r1	7.65	6	2.793E-03	x	
CXCR-4	CXCR-4	dkfz426_3j23r1	1.79	5	7.198E-02	x	
Cyclin B2	CCNB2	dkfz426_73i8r1	1.52	5	7.128E-02	x	
Pax5	Pax5	dkfz426_11e2r1	5.78	4	5.081E-02		168
BASH	BLNK	dkfz426_95h21r1	7.42	6	6.709E-02		88
BAFF	TNFSF13B	dkfz426_69d19r1	2.12	6	6.273E-02		78
ChB6	Gga.37774	dkfz426_33d14r1	1.86	6	5.957E-02		191

The avian DT40 cell line appears to be especially useful for detailed studies of molecular mechanisms in many fields, most notably in the study of the B-cell receptor signalling (Kurosaki 2002), DNA repair and recombination mechanisms (Arakawa et al. 2002; Blagodatski et al. 2009) as well as the genetic regulatory network underlying the B cell to plasma cell transition (I). At the time the DT40 cell line system had not been used as a model for B-cell differentiation (III). This study (Koskela et al. 2003) also laid the foundation for gene targeting studies using the DT40 bursal B cells as a model system for B-cell biology (I; IV) by establishing that the DT40 cell line was most similar to bursal B cells in general and thus retained features of the organ it was derived from (Neiman et al. 1988). The gene expression in bursal B cells was also

compared to the harderian gland plasma cells (Table 5.2.), which was instrumental in interpreting the Pax5 DT40 knock-out study.

Table 5.2. Genes changing between Bursal B cells and Harderian gland plasma cells. The LCF refers to the log fold change, # changing refers to the number of spots (out of four) that have a change larger than two-fold and the p-value is a pair-wise t-test p-value between the bursal B cells and plasma cells.

Gene Name	Symbol	Clone_ID	LFC	# Changing	p-value
IgL chain	IGL@	dkfz426_5g3r1	3.94	4	8.14E-04
MKP1	DUSP1	dkfz426_115l5r1	3.81	4	1.45E-04
DEAD box RNA helicase	DDX5	dkfz426_90d4r1	2.69	4	6.77E-05
MDA-9	SDCBP	dkfz426_97l2r1	2.25	4	1.10E-03
ATF4	ATF4	dkfz426_83k15r1	2.22	4	2.19E-02
IP3KA	ITPKA	dkfz426_89c16r1	2.10	4	6.35E-05
NEDL3	NEDD4-2C6	dkfz426_73f21r1	1.70	4	2.90E-03
Stathmin	STMN1	dkfz426_37l2r1	-4.28	4	2.29E-05
Importin subunit alpha-2	KPNA2	dkfz426_121h2r1	-3.59	4	2.53E-03
B6.3 protein (Bu-1)	Gga.42308	dkfz426_33d14r1	-2.77	4	9.23E-04
CYCB2	CCNB2	dkfz426_73i8r1	-2.37	4	5.22E-04
CXCR4	CXCR4	dkfz426_3j23r1	-2.12	4	4.00E-05
Polymerase lambda	POLL	dkfz426_47b19r1	-2.06	4	1.82E-05
chSNF7 OR VPS32	CHMP4B	dkfz426_75f7r1	-1.93	4	7.42E-04
PRDX1	TDPX2	dkfz426_60k16r1	-1.68	3	6.30E-03
SWAP70	SWAP70	dkfz426_90b17r1	-1.40	4	1.13E-05
H2AFZ	H2AZ	dkfz426_49l13r1	-1.19	4	3.79E-04

5.4. Pax5 DT40 knockout cells undergo the plasma cell transition (I)

The paired box transcription factor 5 (PAX5) was shown to be critical for the maintenance of B cell fate (I). This is in accordance with earlier and parallel studies (Mikkola et al. 2002; Rolink et al. 1999; Delogu et al. 2006). The Pax5 activated genes have been studied in detail (I; Schebesta et al. 2007). However, the most novel result to emerge from this study was that removal of the Pax5 gene in DT40 cells resulted in these cells transitioning to the plasma cell fate (I). The first indications of the plasma cell transition were obtained from the gene expression analysis of the knockout cells (Table 5.3). An expression signature of plasma cells in the avian system was established earlier (Koskela et al. 2003). Genes that were part of this signature were noticed to be changing in a manner that indicated plasma cell behaviour.

The Bursa EST array (I) did not, however, contain any plasma cell specific genes. In order to establish whether the cells had transitioned to the plasma cell fate we showed that the Pax5 -/- DT40 cell not only expressed the plasma cell transcription factor Blimp1 but also repressed the Blimp1 inhibiting transcription factor BCL6, which is expressed at high level by germinal centre B cells. The Pax5 knockout cells also expressed the XBP-1 gene that is required for IgM secretion as well as its secretory splice isoform (Table 5.3).

Accordingly the Pax5 -/- DT40 cell line also initiated the secretion of IgM, indicating that they have fully transitioned to the plasma cell fate. An important indicator that the transition was not a clonal artefact was also the reversion of the terminally differentiated plasma cells back to B cells by over-expression of the Pax5 gene (I).

Table 5.3. The expression levels obtained from Bursa EST array analysis or by quantitative PCR (LightCycler ™, Roche, Germany), relative to the DT40 wild-type cells (II). The results are expressed as fold differences. The n.d. indicates that the experiment was not done. The Pax5 -/- indicates Pax5 knock-out DT40 cells and the Pax5 -/- + Pax5 indicates knock-out cells that have had the Pax5 gene over-expressed to reconstitute the wild-type phenotype.

	Pa5-/-		**Pax5-/- + Pax5**
Gene	**Array**	**Q-RT-PCR**	**Q-PCR**
CD79b/Igb	0.4	0.06	1.07
BLNK	0.53	0.39	1.38
Lyn	0.33	0.16	0.94
Btk		0.5	n.d.
VpreB3	0.18	0.03	0.89
IgL		2.55	n.d.
HSP70	5.8	9.7	n.d.
ITM2A	0.53	0.38	n.d.
HMG-17	0.45	0.45	n.d.
Caspase 3	0.53	0.43	n.d.

5.5. Helios and Ikaros family evolution (II)

Alternative splicing of Ikaros family members has been shown to contribute to lymphoid malignancies (Rebollo et al. 2003, Mullighan et al. 2007). We show in this article that Helios has multiple alternative splicing isoforms which are evolutionarily conserved (Figure 3.). Also the insertion of a novel exon was discovered in the chicken Helios gene. Helios expression also preceded Ikaros in the ontogeny. Expression of Helios in the bursa of Fabricius, germinal centres and B-cell lines suggests a role for Helios also in the B-cell lineage. Expression in the DT40 cell line means that it is possible to study the loss of Helios function in that model (Alinikula et al. 2010).

The evolution of the Ikaros family was also investigated. The four lymphocyte associated members of the Ikaros family: Ikaros (IKFZ1), Helios (IKZF2), Aiolos (IKFZ3) and Eos (IKFZ4) form a symmetric tree with Ikaros and Aiolos in one branch and Helios and Eos in the other. This kind of a tree and the association of the Ikaros family with syntenic regions next to the Hox clusters lead us to propose that the Ikaros family duplicated and that all the members were retained in the second genome-wide duplication in accordance with the 2R hypothesis proposed by Susumo Ohno (Ohno, 1970). Since the second genome-wide duplication coincides with the emergence of the adaptive immune system, Ikaros family could have played an important role in that as well. The connection of the four members to the jawless vertebrate Ikaros family members was investigated. This is interesting because the jawless vertebrates have not

undergone the second genome-wide duplication event. They also lack the adaptive immune system as it is known in jawed vertebrates, although they have developed a similar system through convergent evolution (Cooper & Alder 2006, Alder et al. 2008).

5.6. Meta-analysis of the Pax5 regulated genes and pathways (III, IV)

Combining the mouse and chicken Pax5 knock-out gene expression results makes it possible to establish a core set of evolutionarily conserved and Pax5 regulated genes. The chicken gene expression studies and results of the DT40 cell line knockouts were compared to mammalian gene expression studies and results of the mouse knockouts (III, IV).

Table 5.4 Genes that are down-regulated in the cross-species meta-analysis. The top 50 results are shown, all of which have a very low multiple testing corrected significance level (FDR q.value = 0). The Ch.DN and Ch.UP indicate that in the species specific compendium the gene is differentially expressed either being reduced or increased, respectively (FDR q.value <0.05); Likewise for the mouse (Mm).

Name	*RP/Rsum*	*logFC*	*q.value*	*Ch.DN*	*Ch.UP*	*Mm.DN*	*Mm.UP*
BLNK	2.31	-2.55	0	**x**		**x**	
VPREB3	4.52	NA	0	**x**		**x**	
BTG1	29.52	-1.16	0	**x**		**x**	
BCAT1	36.32	-1.09	0	**x**		**x**	
ERO1LB	41.19	NA	0	**x**		**x**	
CD79B	45.66	NA	0	**x**		**x**	
FOXO1	45.69	-1.67	0	**x**		**x**	
SNN	61.53	NA	0			**x**	
AP1AR	67.00	-1.28	0	**x**		**x**	
DCBLD1	69.75	NA	0	**x**		**x**	
SNX2	74.95	-0.68	0			**x**	
CCNG2	84.49	NA	0	**x**		**x**	
LYN	88.29	NA	0	**x**		**x**	
RRM2	95.94	-0.55	0			**x**	
PLEKHA2	101.11	-0.92	0	**x**		**x**	
CTH	109.52	-0.82	0			**x**	
IFI30	111.57	-0.79	0	**x**		**x**	
NUAK1	119.85	NA	0	**x**		**x**	
IRF8	121.87	-0.86	0	**x**		**x**	
MYH10	122.69	-0.65	0			**x**	
DPYSL2	127.37	NA	0	**x**		**x**	
CCNA2	130.44	-0.52	0			**x**	
GPD1L	131.76	-0.68	0			**x**	
BNIP3	135.19	-1.04	0	**x**		**x**	
GPSM2	137.93	-0.79	0	**x**		**x**	
PRKCB	141.22	NA	0	**x**		**x**	
NEIL3	141.59	-0.44	0			**x**	
EZH2	141.99	-0.46	0			**x**	

Name	*RP/Rsum*	*logFC*	*q.value*	*Ch.DN*	*Ch.UP*	*Mm.DN*	*Mm.UP*
NUSAP1	143.37	NA	0	**x**		**x**	
SMTN	147.18	-0.45	0			**x**	
SERINC5	148.13	-0.52	0			**x**	
LXN	150.99	NA	0			**x**	
EIF2AK3	160.18	NA	0			**x**	
APITD1	167.71	-0.71	0			**x**	
GLDC	171.55	-0.68	0			**x**	
FBXL12	171.67	-0.40	0			**x**	
GLCCI1	172.48	-0.59	0			**x**	
SBK1	175.39	-0.46	0			**x**	
FMR1	177.91	-0.58	0			**x**	
FBXO5	184.10	-0.49	0			**x**	
RRAS2	184.65	-0.66	0	**x**		**x**	
GTF2A2	189.08	-0.34	0			**x**	
MKI67	190.85	-0.64	0	**x**		**x**	
NDC80	194.23	-0.43	0			**x**	
CCNB2	194.28	NA	0			**x**	
NFKBIA	196.28	-0.84	0	**x**		**x**	
ANKRD33B	196.46	-0.37	0			**x**	
EIF4E3	201.41	-0.40	0			**x**	
IPMK	202.62	-0.40	0			**x**	
CSRP2	210.88	NA	0			**x**	

Table 5.5 Genes that are up-regulated in the cross-species meta-analysis. The top 50 results are shown, all of which have a very low multiple testing corrected significance level (FDR q.value = 0). The Ch.DN and Ch.UP indicate that in the species specific compendium the gene is differentially expressed and either being reduced or increased, respectively (FDR q.value <0.05); Likewise for the mouse (Mm).

Name	*RP/Rsum*	*logFC*	*q.value*	*Ch.DN*	*Ch.UP*	*Mm.DN*	*Mm.UP*
FRMD4B	29.92	0.74	0				**x**
SERPINB1A	32.02	NA	0				**x**
SLC25A1	42.06	0.65	0				**x**
GYG	51.08	1.02	0		**x**		**x**
PARP1	75.67	0.50	0				**x**
CASP6	79.80	NA	0				**x**
PIK3R1	83.22	0.49	0				**x**
WAPAL	86.95	0.52	0				**x**
PRKAR2A	87.63	NA	0				**x**
DCTD	93.33	0.70	0				**x**
MYC	96.52	0.58	0				**x**
DTX4	96.96	NA	0				**x**
PRDX6	100.60	0.40	0				**x**
TMEM66	103.01	0.46	0				**x**
TNFSF11	107.52	NA	0				**x**
PECR	111.39	0.51	0				**x**
RNF130	119.76	0.25	0				**x**
C230081A13RIK	127.52	NA	0				**x**
ST5	139.01	0.49	0				**x**

Name	*RP/Rsum*	*logFC*	*q.value*	*Ch.DN*	*Ch.UP*	*Mm.DN*	*Mm.UP*
DAAM1	140.68	NA	0				**x**
EHD3	143.41	NA	0				**x**
NCOA4	151.88	NA	0				**x**
ZFHX3	151.94	NA	0				**x**
SATB1	154.70	0.27	0				**x**
DNMT3B	157.92	0.49	0				**x**
TULP4	160.14	NA	0				**x**
PPP1R3B	163.90	0.44	0				**x**
EMP1	167.87	NA	0				**x**
PCYT1A	173.54	0.80	0		**x**		**x**
CD28	181.42	NA	0				**x**
SPP1	184.93	0.44	0				**x**
TEX2	193.08	NA	0				**x**
HNRPLL	193.81	0.56	0				**x**
KLHDC2	194.76	NA	0				**x**
NCLN	199.04	NA	0				**x**
IQGAP2	199.23	0.60	0				**x**
TMSB10	203.39	NA	0				**x**
ELK3	205.02	NA	0				**x**
SESN1	205.12	NA	0		**x**		
XRCC5	209.43	0.35	0				**x**
GNL3	209.51	NA	0				**x**
FTH1	211.02	0.46	0				**x**
RHOQ	217.13	NA	0		**x**		
EPCAM	218.92	0.44	0				**x**
SLC2A3	220.17	0.51	0				**x**
ENDOD1	226.29	0.36	0				**x**
ABI2	226.30	0.42	0				**x**
LEPROTL1	229.19	NA	0				**x**
ATAD3A	234.92	NA	0		**x**		
ARHGAP18	251.48	0.28	0				**x**

The comparisons between knock-out studies (IV), especially the down regulated genes, were largely concordant between chicken and mammals. Several known Pax5 regulated genes such as BLNK and CD79B were regulated in the same way in both species.

The differentially expressed sets from the Cancer gene neighbourhood class can be grouped around a few sets: sets containing cell cycle genes, related to the GNF2_BUB1 gene set; GNF2_LYN related sets containing B-cell receptor signalling genes. The third group is the GNF2_ANP32B gene set that is left out as not being related either to the LYN or BUB1 sets. The upregulated gene sets are all related to GNF2_CD7 or GNF2_IL2RB sets.

Table 5.6 Cancer Gene Neighbourhood (CGN) gene sets from the MsigDB which are differentially expressed in the cross-species meta-analysis (FDR q.value <0.05). The GNF2_BUB1, GNF2_LYN, GNF2_ANP32B and GNF2_CD7 related sets are indicated on the right.

SetName	*Type*	*RP/Rsum*	*Score*	*q.value*	*BUB1*	*LYN*	*ANP32B*	*CD7*
GNF2_MKI67	*down*	2.67	-1.14	0	**x**			
GNF2_CDC2	*down*	3.72	-1.14	0	**x**			
GNF2_H2AFX	*down*	4.64	-1.07	0	**x**			
GNF2_CENPE	*down*	5.87	-1.09	0	**x**			
GNF2_ESPL1	*down*	6.85	-1.00	0	**x**			
GNF2_CENPF	*down*	7.80	-1.06	0	**x**			
GNF2_CCNB2	*down*	9.57	-0.99	0.0029	**x**			
GNF2_CCNA2	*down*	10.02	-0.97	0.0038	**x**			
GNF2_HMMR	*down*	10.93	-0.91	0.0033	**x**			
GNF2_RRM2	*down*	10.95	-0.95	0.003	**x**			
GNF2_SMC2L1	*down*	11.15	-0.96	0.0027	**x**			
GNF2_PCNA	*down*	12.82	-0.91	0.0067	**x**		**x**	
GNF2_CDC20	*down*	14.53	-0.87	0.0085	**x**			
GNF2_CKS2	*down*	14.56	-0.86	0.0079	**x**			
GNF2_LYN	*down*	17.93	-0.74	0.012		**x**		
GNF2_BUB1	*down*	19.80	-0.74	0.0213	**x**			
GNF2_TTK	*down*	20.11	-0.73	0.0218	**x**			
GNF2_BUB1B	*down*	22.56	-0.67	0.0306	**x**			
GNF2_FEN1	*down*	22.67	-0.64	0.0295	**x**		**x**	
GNF2_MCM4	*down*	23.22	-0.65	0.0335	**x**			
GNF2_RRM1	*down*	24.93	-0.61	0.0452	**x**		**x**	
GNF2_SMC4L1	*down*	26.22	-0.61	0.0505	**x**		**x**	
GNF2_CKS1B	*down*	26.95	-0.60	0.0504	**x**			
GNF2_RFC3	*down*	28.55	-0.56	0.06	**x**			
GNF2_RFC4	*down*	32.93	-0.51	0.092	**x**			
GNF2_ANP32B	*down*	34.38	-0.44	0.0988			**x**	
GNF2_CD48	*down*	34.77	-0.72	0.0978		**x**		
GNF2_RAB7L1	*up*	4.59	0.79	0				**x**
GNF2_CD7	*up*	10.41	0.59	0.03				**x**
GNF2_IL2RB	*up*	10.78	0.56	0.02				**x**
GNF2_SNRK	*up*	14.84	0.44	0.05				**x**
GNF2_CASP8	*up*	16.81	0.26	0.048				**x**

The generally most easily interpretable gene sets are canonical pathways and other curated sets that contain genes belonging to a specific pathway or are derived from a certain experiment (Subramanian et al. 2005). The meta-analysis of these sets brought up sets that were not otherwise seen as being differentially expressed in both of the species.

Table 5.7. Curated gene sets from the MsigDB. A. Down-regulated gene sets at FDR q.value<0.05. B. Up-regulated gene sets at FDR q.value < 0.05. The GSA overall, **C**hicken (Ch) and **M**ouse (Mm) scores are shown separately. Results from the organism-specific compendia are shown for each organism as well and whether they pass FDR <0.5 threshold for significance.

A.

Name	RP/Rsum	Score	C.Score	M.Score	Q.value	P.value	Ch.UP	Ch.DN	Mm.UP	Mm.DN
LAMB_CYCLIN_D3_GLOCUS	6.93	-0.90	-0.31	-1.93	0	0				x
KLEIN_PEL_DN	15.80	-0.84	-0.63	-1.24	0	0		x		x
DOX_RESIST_GASTRIC_UP	17.94	-0.73	-0.24	-1.67	0	0				x
ZHAN_MM_CD138_PR_VS_REST	19.89	-0.63	-0.09	-1.35	0	0				x
IRITANI_ADPROX_UP	22.05	-0.72	-0.11	-1.92	0	0				x
P21_P53_MIDDLE_DN	24.28	-0.74	-0.53	-1.17	0.0012	0				
BCRPATHWAY	28.58	-0.75	-0.73	-0.55	0.0033	0		x		x
GREENBAUM_E2A_UP	29.21	-0.62	-0.29	-1.36	0.003	0				
LEE_TCELLS3_UP	33.94	-0.58	-0.20	-1.12	0.0036	0				
DORSEY_DOXYCYCLINE_UP	34.69	-0.56	-0.08	-0.80	0.0042	0				
HG_PROGERIA_DN	38.74	-0.57	-0.20	-1.32	0.0062	1.00E-04				x
INSULIN_ADIP_INSENS_DN	40.43	-0.62	-0.39	-1.00	0.0071	1.00E-04				x
BASSO_GERMINAL_CENTER_CD40_DN	43.62	-0.50	-0.17	-0.63	0.0087	1.00E-04				x
CROONQUIST_IL6_STARVE_UP	44.29	-0.59	-0.39	-1.10	0.0088	1.00E-04				
P21_P53_ANY_DN	44.59	-0.59	-0.41	-0.94	0.0083	1.00E-04				
GOLDRATH_CELLCYCLE	44.60	-0.62	-0.50	-1.12	0.0079	1.00E-04				
OLDAGE_DN	49.42	-0.49	-0.15	-1.14	0.0105	1.00E-04				x
BREAST_DUCTAL_CARCINOMA_GENES	50.78	-0.49	-0.12	-0.98	0.0105	1.00E-04				
IDX_TSA_UP_CLUSTER3	51.37	-0.52	-0.29	-0.95	0.01	1.00E-04				x
VERNELL_PRB_CLSTR1	52.19	-0.46	-0.08	-1.24	0.0096	1.00E-04				x
GOLUB_ALL_VS_AML_UP	55.82	-0.45	-0.13	-0.80	0.0104	2.00E-04				
PARK_HSC_VS_MPP_DN	55.99	-0.54	-0.31	-0.89	0.0104	2.00E-04		x		
INSULIN_ADIP_INSENS_UP	56.68	-0.50	-0.33	-0.77	0.0115	2.00E-04				
CROONQUIST_IL6_RAS_DN	57.42	-0.64	-0.72	-0.84	0.013	2.00E-04				
P21_MIDDLE_DN	61.82	-0.43	-0.05	-1.03	0.0175	3.00E-04				
ADIPOGENESIS_HMSC_CLASS2_UP	61.91	-0.55	-0.27	-1.67	0.0169	3.00E-04				x
H2O2_CSBRESCUED_C1_UP	66.96	-0.47	-0.20	-0.93	0.021	4.00E-04				x
CANCER_UNDIFFERENTIATED_META_UP	67.29	-0.41	-0.06	-0.91	0.0203	4.00E-04				x
AGEING_LYMPH_DN	68.23	-0.37	0.08	-1.18	0.0203	4.00E-04				x
HSA04662_B_CELL_RECEPTOR_SIGNALING_PATHWAY	68.39	-0.61	-0.80	-0.27	0.0197	4.00E-04		x		
SHEPARD_GENES_COMMON_BW_CB_MO	68.61	-0.50	-0.41	-0.77	0.0194	4.00E-04				
LE_MYELIN_UP	69.59	-0.46	-0.27	-0.87	0.0194	5.00E-04				x
P21_ANY_DN	72.11	-0.44	-0.21	-1.00	0.0214	5.00E-04				
SA_B_CELL_RECEPTOR_COMPLEXES	81.14	-0.43	-0.35	-0.19	0.0297	8.00E-04		x		
CITED1_KO_WT_UP	82.17	-0.33	0.09	-1.08	0.0308	8.00E-04				x
SERUM_FIBROBLAST_CELLCYCLE	83.68	-0.42	-0.25	-0.81	0.0317	9.00E-04				
ADIP_DIFF_CLUSTER5	84.36	-0.39	-0.22	-0.70	0.0321	9.00E-04				
ZHAN_MMPC_EARLYVS	84.63	-0.54	-0.66	-0.51	0.0314	9.00E-04		x		
IGF_VS_PDGF_DN	85.13	-0.40	-0.27	-0.59	0.0309	9.00E-04		x		
CMV_IE86_UP	85.74	-0.58	-0.72	-0.56	0.0311	9.00E-04				x
CYTOKINEPATHWAY	89.01	-0.43	-0.08	-1.95	0.0361	0.0011				
SA_MMP_CYTOKINE_CONNECTION	91.68	-0.34	0.08	-1.03	0.0417	0.0013				
ST_B_CELL_ANTIGEN_RECEPTOR	94.24	-0.53	-0.70	-0.19	0.0469	0.0015		x		
TNF_AND_FAS_NETWORK	94.54	-0.47	-0.47	-0.52	0.0467	0.0015				
RADIATION_SENSITIVITY	95.82	-0.37	-0.17	-0.87	0.0484	0.0016				x
ZHAN_TONSIL_PCBC	96.09	-0.47	-0.55	-0.47	0.0478	0.0016		x		
YU_CMYC_DN	97.14	-0.37	-0.25	-0.61	0.049	0.0017				

B.

Name	RP/Rsum	Score	C.Score	M.Score	Q.value	P.value	Ch.UP	Ch.DN	Mm.UP	Mm.DN
MITOCHONDRIAL_FATTY_ACID_BETAOXIDATION	11.04	0.69	0.14	1.56	0	0				
MUSCLE_MYOSIN	16.12	0.78	0.22	1.48	0.0033	0				
LEE_TCELLS5_UP	19.59	0.59	0.11	1.26	0.0025	0				
HBX_HEP_UP	24.50	0.67	0.60	0.85	0.002	0	x			
N_GLYCAN_BIOSYNTHESIS	47.56	0.64	0.74	0.78	0.015	1.00E-04	x			
PROPANOATE_METABOLISM	52.34	0.44	0.23	0.77	0.0175	1.00E-04			x	
ZELLER_MYC_UP	56.80	0.70	0.97	0.54	0.0193	2.00E-04	x			
FATTY_ACID_DEGRADATION	60.73	0.41	0.17	0.76	0.0213	2.00E-04				
VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	62.80	0.56	0.66	0.58	0.0241	3.00E-04	x			
ARFPATHWAY	63.23	0.43	0.36	0.76	0.025	3.00E-04	x			
IL7PATHWAY	67.70	0.48	0.43	0.69	0.0286	4.00E-04	x			
CITRATE_CYCLE_TCA_CYCLE	68.05	0.42	0.27	0.69	0.0277	4.00E-04				
HSA00440_AMINOPHOSPHONATE_METABOLISM	70.56	0.71	1.15	0.24	0.0296	5.00E-04	x			
HSA00020_CITRATE_CYCLE	76.10	0.44	0.38	0.63	0.035	6.00E-04				
BENNETT_SLE_UP	81.10	0.55	0.53	0.63	0.0425	8.00E-04				
HSA00640_PROPANOATE_METABOLISM	81.81	0.37	0.23	0.82	0.0424	8.00E-04			x	

6. DISCUSSION

6.1. Genetic regulatory networks in the B-cell to plasma cell transition

6.1.1. Meta-analysis enables comparison of diverse data sets

Biomedical sciences are among the ones that are most keenly influenced and transformed by technological innovation. The advent of whole-genome sequencing has transformed biochemistry and molecular biology from a single-gene reductionist approach into a more holistic systems-wide approach. Microarrays have enabled researchers to measure the expression of all the known genes in the genome (Allison et al. 2006).

Gene knock-out studies are difficult to perform on animal cells or on whole animals. However, notable exceptions include the chicken DT40 B cell line that readily permits deletion of any expressed gene. RNAi methods utilizing short pieces of double-stranded RNA to degrade transcripts can also be used to knock down or to reduce the expression levels of genes (Boutros & Ahringer 2008). Both approaches to gene silencing can be very conducive to microarray studies of gene function (III). Small molecular or protein based drugs can also be used to manipulate gene expression (Lamb et al. 2006; Lamb 2007).

Over expression or ectopic expression studies can be performed relatively easily but the results are artificial in nature, either because the level of expression of the gene is too high or because the biological context does not resemble the normal *in vivo* situation. In contrast, the DT40 cell line model permits complete removal of a gene, such as the Pax5 transcription factor (I). Hence, residual levels that might be left after RNAi treatment do not interfere with the comparison of knock-out to wild-type cells (I, IV). Mouse whole organism gene knock-out studies can be performed and are highly informative. However, while a whole organism is generally preferable to a cell line as the model, cells inside an organism are better able to adapt to the removal of a key transcription factor, for instance by selectively expanding a viable sub-population of cells. Despite their limitations, DT40 cell line models can thus give information that is complementary to the mouse knock-out models. The results of the various knock-out studies can also be directly compared to derive new information on broadly conserved regulatory mechanisms (IV).

One could argue that mRNA levels inside the cell reflect the ongoing gene regulatory programs that are active. Indeed, functional signatures, biochemical pathways or gene expression modules that change in response to cell differentiation or treatments are useful tools for data reduction (Ollila & Vihinen 2007). Thus, gene class testing facilitates the understanding of the ongoing biological processes (Ashburner et al. 2000, Subramanian et al. 2005). However, as the number of comparisons increases, it becomes increasingly difficult to compare them to each other.

Meta-analysis approaches can be used to find consistently regulated genes across different studies. Compared to Venn diagrams, which are usually used for comparing different experiment with each other, meta-analysis methods have an advantage in that they are not threshold-dependent. Definitions of thresholds can vary substantially and may be difficult to standardize across experiments. In most meta-analysis approaches a p-value across experiments is calculated (Troyanskaya 2005), but only at the end, meaning that each dataset is treated the same. Therefore, when three or more experiments are compared to each other, meta-analysis approaches could be used instead of Venn diagrams, with probably better results.

It has been argued that results of the enrichment analysis of gene sets are more sensitive and reproduce better across different microarray platforms than the differential expression results of individual genes (Nam et al. 2008, Manoli et al. 2006). Gene class testing results, done with GSEA-types of methods, can also be combined to find out which classes are most consistently up or down regulated across species, different platforms or across heterogeneous experiments (IV).

6.1.2. The central role of Pax5 in the B-cell to plasma cell transition

The results we obtained regarding the transition of Pax5 -/- B DT40 cells into the plasma cell fate (I) were unexpected and surprising at the time of the finding. This was because it had been shown that in mice the Pax5 gene knockout early in B-cell differentiation leads to de-differentiation of the B cells and their reversion to a lymphoid progenitor cell-like state. The Pax5 -/- pro-B cells can reconstitute the entire lymphoid lineage, except for the B cells (Mikkola et al. 2002; Rolink et al. 1999).

On the plasma cell side, the Blimp1 factor has a central role in plasma cell differentiation. The Blimp1 mouse knock-out cells lack plasma cells, and over-expression of Blimp1 in germinal centre B cells can induce plasma cell differentiation (Turner et al. 1994; Schliephake & Schimpl 1996; Shaffer et al. 2002). Indeed, the induction of the Blimp1 transcription factor expression is central for plasma cell differentiation (Shapiro-Shelef & Calame 2005; Schmidlin et al. 2009). However, the steps leading to Blimp1 induction have been under debate. Previously, the induction of the Blimp1 protein was seen as the initial step in plasma cell differentiation (Shapiro-Shelef & Calame 2005).

The idea that the repression of the B cell program, starting with Pax5, is the initial step in plasma cell differentiation has by now gained significant traction (Kallies et al. 2007; Klein & Dalla-Favera 2007), corroborating and expanding the findings in Article I. Immune deficient Rag1 -/- mice reconstituted with cells expressing inactive Blimp1 protein (Kallies et al. 2007) managed nevertheless to produce small but detectable amounts of all immunoglobulin subtypes. This stage, which was reached independently of Blimp1 induction, was termed the pre-blasmablast stage. It is characterized by reduced Pax5 expression, downregulation of the B-cell specific genes as well as low levels of immunoglobulin secretion. This again suggests that the downregulation of Pax5 expression is the initial stage of the plasma cell differentiation process.

6.1.3. The anatomy of the B-cell to plasma cell genetic switch

The structure of the Prdm1 gene promoter explains some of the observations of the previous knock-out studies (Figure 9). The upregulation of the Blimp1 protein when the *Pax5* gene is deleted (I) is partly explained by the presence of the Pax5 binding site on the Prdm1 gene promoter (Mora-López et al. 2007). Indeed, it seems that a whole host of factors, including at least Pax5, BCL6 and Bach2, need to be taken out of action before Blimp1 induction can proceed (Calame 2008; De Vos et al. 2006). The IRF4 gene is required for the activation of Blimp1 (Kwon et al. 2009; Calame 2008; Klein et al. 2006; Fillatreau 2006) (Figure 9).

The signals that the B cell needs in order to make a decision on plasma cell differentiation seem to be mainly integrated at the *Pdrm1* gene promoter (Calame 2008; Saito et al. 2007). Several mechanisms are known for the downregulation of the BCL6 gene that seem to be directly related to B-cell activation (Calame 2008). Activation of the B-cell receptor (BCR) can repress BCL6 via the Akt signalling pathway (Calame 2008). The CD40 co-activation signal acts via the NF-κB by activating the IRF4 gene (Saito et al. 2007). This is doubly effective because IRF4 not only represses BCL6 but activates the *Prdm1* gene as well (Calame 2008; Saito et al. 2007; Klein et al. 2006).

The signalling cascades that initiate Pax5 dowregulation have remained unknown (Calame 2008). Recently, a regulatory connection between the antigen receptor activation and the E2A protein was proposed (Hauser et al. 2009). The levels of Pax5, BCL6, MITF, Ets-1, Fli-1, and Spi-B are all reduced rapidly upon antigen receptor activation; much of the reduction in their mRNA levels occurred within 30 minutes. The signal is mediated using Ca2+ as the second messenger and calmodulin (CaM) binding to the E2A protein (Hauser et al. 2009). The rapidity of the response indicates direct regulatory interactions but on the other hand the E2A protein has also been shown to be dispensable for the maintenance of the B cell program and for the plasma cell differentiation, although it potentiates germinal centre B cell survival (Kwon et al. 2008). Further mechanistic studies are likely to clarify this point.

EBF1, in addition to Pax5, is required for B lineage commitment (Ramirez et al. 2010). However, although Pax5 and EBF1 regulate the B cell program together (Treiber et al. 2010), there is no evidence of EBF1 being mediator of the B cell to plasma cell transition.

The various types and subsets of B cells, including naïve B cells, germinal centre B cells, memory B cells and pre-blasmablasts, take somewhat different routes to Blimp1 repression (Schmidlin et al. 2009) (Figure 10). Memory B cells have not downregulated Pax5, at least not to the same extent, but do reduce or eliminate BCL6 expression (Schmidlin et al. 2009; Calame 2008). Plasma cells and the stimuli they are generated from also differ (Fairfax et al. 2008; Schmidlin et al. 2009; Calame et al. 2008). Part of the complexity of the regulatory network underlying plasma cell differentiation (Figures 9 and 10) may therefore stem from the need to integrate different stimuli in various cell-type subsets and at different times.

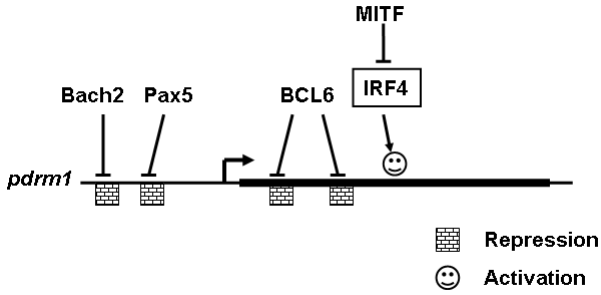

Figure 9. Regulation of the *prdm1* gene proximal promoter indicates that repression by B cell and germinal cell factors needs to be lifted before induction.

6.1.4. Logic of the switch from a network biology perspective

The Pax5 and Blimp1 proteins conform to a double-negative auto-regulatory feedback loop (Figure 10; Figure 4.) (I; Mora-López et al. 2007; Schmidlin et al. 2008; Schmidlin et al. 2009; Rothenberg 2007). We speculated at the time that Blimp1 was upregulated as a consequence of low BCL6 levels. However, since the Prdm1 gene contains a functioning binding site for the Pax5 protein (Mora-López et al. 2007), the direct regulatory model is more likely. Further confirmation in a mammalian system is provided by the upregulation of the Blimp1 transcipt in mature B cells that are made Pax5 deficient (Delogu et al. 2006). The B-cell lineage associated genes as well as the Pax5 gene itself are in turn repressed by the Blimp1 protein (Lin et al. 2002; Shaffer et al. 2002). The BCL6 protein also represses the Prdm1 gene with the help of the MTA3 cofactor (Fujita et al. 2004) and is itself repressed by it (Shaffer et al. 2002). Interestingly, IRF4 and Blimp1 can also induce each other (Lu 2008; Kwon et al. 2009). This second but positive auto-regulatory loop is likely to potentiate Blimp1 induction in the face of repression by the B cell factors (Lu 2008). Feedback loops in general thus seem to be a recurring theme in developmental genetic circuits (Rothenberg 2007; Alon 2007 98-104).

Robustness is also a necessary property of biological regulatory circuits (Alon 2007). The concentrations of cellular partners are small and sometimes only a few dozen protein macromolecules may be present inside a cell, and the genes are present in only two copies. This creates stochastic variation in cellular functions.

The double negative circuit seems to be vulnerable to stochastic variations. However, it may be that the multi-step de-activation of *Prdm1* gene repression confers robustness to the transition. Activation and terminal differentiation are less likely to happen by accident if the lifting of any one repressor from the promoter is not sufficient to activate it. This is perhaps similar to the mechanism termed kinetic proofreading that is observed in the activation of the B-cell receptor (Alon 2007, 182-188). The need to phosphorylate several ITAM motifs during receptor activation forces the antigen to remain bound to the receptor for a longer time, which helps to ensure sufficient specificity. Also second signals in the form of co-activators, such as CD40, are almost

always required for activation. These signals are integrated and interpreted at the *Prdm1* gene promoter.

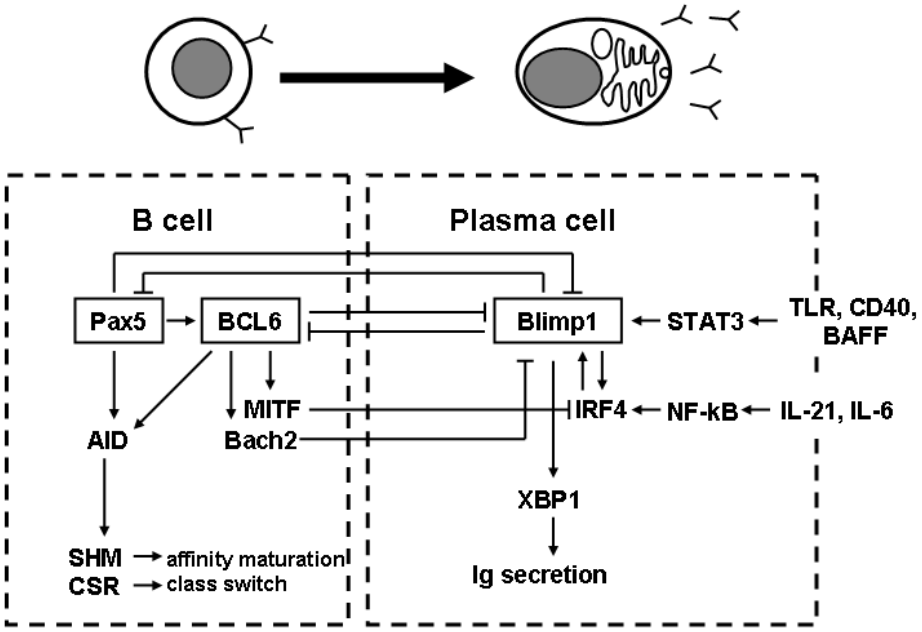

Figure 10. Transcriptional regulatory switch for the B cell to plasma cell transition has an overall double negative feedback loop structure between the B cell program and the plasma cell program.

6.1.5. New technologies to study developmental genetic switches

The DT40 cell line system excels in mechanistic studies (I, III, IV). Further avenues of research include generating knockouts in the DT40 cell line system of the other players in the B-cell to plasma cell regulatory network, and observing the effects on differentiation. These include BCL6, IRF4/8 as well as MITF and perhaps others. In order to elucidate further their roles in the regulatory network, it would be desirable to carry out ChIP-seq as well as gene expression analyses (Treiber et al. 2010). Studying Pax5 in this manner in the DT40 system would also be advantageous. Clarifying the set of transcription factors that are predicted to regulate the same set of genes with Chip-seq would also provide information on the clustering of the transcription factor binding sites on promoters or enhancers (Treiber et al. 2010). Such information has not yet been readily available due to the small number of genome-wide ChIP studies carried out so far (Farnham 2009).

Integration of the gene expression and Chip-seq data can further define the regulatory relations in the network. The target sites and genes identified by Chip-seq and gene expression can be broadly divided into three categories. 1) Genes whose expression is changed and which have a proximal promoter or enhancer binding site for the transcription factor; 2) genes whose expression is changed but there is no binding observed and 3) genes with binding but no transcriptional change. The direct targets

bind the regulatory factor, whereas the indirect targets do not. The third category may represent binding that requires other factors to have an effect, maybe in another cell type or under different conditions. The presence of biological noise also cannot be ruled out (Visel et al. 2009; Park 2009; Farnham 2009).

The Pax5 transcription factor can act both to activate as well as repress genes. But how is this behaviour regulated or how does it come about? Some features of the regulatory regions, at either the chromatin or most likely the DNA sequence level, are likely to decide whether a gene is activated or repressed by Pax5. One answer seems to be that Pax5 co-operates with various partners in gene activation (Schmidlin et al. 2009). These include Ets factors; Runx1 and PU.1. On the other hand, various repressive transcription factors, including MITF and BCL6, bind the *Prdm1* gene promoter alongside Pax5. But it is thought that these factors act independently and there has been no indications that they would define whether Pax5 acts to repress or activate the *Prdm1* gene (Calame 2008).

Post-translational modifications of the Pax5 protein might also define whether it acts as an activator or a repressor, by changing the co-factors that are able to bind to it. Groucho-family co-factors mediate repressive actions of Pax5 (Eberhard et al. 2000). However, unless post-translational modifications change the Pax5 binding specificity, they still do not explain how the Pax5 protein is able to simultaneously carry out both activating and repressing regulatory actions.

Genome-wide analysis of Pax5-binding regulatory region might also make it possible to determine if there are common sequence features that are associated with promoters or enhancers that are either repressed or activated by Pax5. Sequence motifs that are closely associated with the Pax5 motif might point to transcription factors that interact with the Pax5 protein to flip the factor from an activator to a repressor, or they might preferentially bind either type of complexes. In some cases PU.1 co-recruitment can specify repressive actions (Linderson et al. 2004). It may also be that repression is the default mode of action of Pax5. In that case one might try to assess whether the repressive sites have motifs that are closer to the consensus sequence, and might therefore have higher affinity for the Pax5 factor; or maybe these sites contain more than one Pax5 binding site.

A recent study of the early B cell factor 1 (EBF1) gene (Treiber et al. 2010) combines genome-wide Chip-seq analysis with gain- and loss-of-function gene expression analyses. A third of Pax5 targets in early B cells are also bound by EBF1, indicating that these factors regulate their targets together and demonstrating a way to build a network for Pax5 actions at the B cell to plasma cell transition as well.

In short, by enumerating binary relationships between regulators and the regulatory regions of target genes, one can generate a network of regulatory relationships inferred from the ChIP-seq data and the expression data after perturbations of the regulators. This can be visualized and analyzed with the Cytoscape software (Aittokallio & Schwikowski 2006; Cline et al. 2007). Furthermore, it might be possible to infer regulatory programs automatically using methods such as those implemented by the Allegro software (Halperin et al. 2009). Such automatically-inferred networks could

then be compared to networks derived from literature reviews. Unbiased regulatory network inference may also uncover unexpected relationships.

6.1.6. Suggestions for modelling of the B-cell to Plasma cell transition

Modelling is emerging as a complement to empirical studies to promote the understanding of transcriptional regulatory circuits (Kim et al. 2009; Karlebach & Shamir 2008). The discussion of cellular states in terms of the expression of transcriptional regulators seems to pre-suppose that it should be possible to predict the expression levels of at least the majority of the genes in the cell based on the levels of these regulators. The complexity of the gene regulatory network (Figure 10) underlying the B cell to plasma cell transition suggests that modelling it quantitatively would help to clarify the dynamics of its behaviour and for instance help to set thresholds for how much stimulus is needed to push the network from one state to another (Alon 2007). Modelling frameworks, such as the Kappa language implemented by the Cellucidate.com, are being developed that are optimized for tackling biological problems (Feret et al. 2009; Webster 2009).

Better understanding of the dynamics of the regulatory actions would help to define parameters for modelling. The chicken DT40 system may prove to be useful in this by being more amenable to mechanistic studies, such as GFP reporter knock-ins into a gene locus to monitor its activity (Kwon et al. 2008).

6.2. Evolutionary aspects of gene regulation

6.2.1. Conservation of gene expression patterns and regulatory programs

Gene regulatory networks are broadly conserved between organisms (I; III; IV; Erwin & Davidson 2009; Xie et al. 2005). Evolutionary studies can also be carried out using gene-expression patterns (Ettwiller et al. 2008; Ramialison et al. 2008; Chan et al. 2009). Therefore, a comparative evolutionary perspective can be useful for studies of regulatory networks as well (IV, Xie et al. 2005).

Meta-analysis approaches (Hong et al. 2006) can be used to find consistently regulated genes across different conditions which are likely to be the most pertinent targets for regulation, not biological noise (III; IV; Rasche et al. 2008; Alles et al. 2009). The combination of expression results from different species can help to focus on the most important genes (III; IV; Sun et al. 2007; Sweet-Cordero et al. 2005). Highly conserved core parts of the network (Chan et al. 2009; Erwin & Davidson 2009) could then be identified through comparative gene expression analysis or comparative transcriptomics (Zhou & Gibson 2004) – which, in analogy to phylogenetics, could be referred to as phylotranscriptomics (III).

6.2.2. Cross-species meta-analysis of the Pax5 program

A cross-species meta-analysis of the Pax5 regulated expression programs both at the gene and pathway levels was undertaken in order to identify genes that are Pax5 regulated in mouse and chicken B cells (I; IV; Schebesta et al. 2007; Nutt et al. 1998; Pridans et al. 2008). The pathway enrichment analysis employed the GSA R package (Efron & Tibshirani 2006), and combining results from different species identifies pathways whose Pax5 regulation is evolutionarily conserved.

Several genes representing the core of the B cell gene expression program, and genes down-regulated after the Pax5 knock-out in both mouse and the chicken (I; Schebesta et al. 2007; Pridans et al. 2008), are similarly regulated across species. The top-most genes affected in both organisms also tended to be similar: BLNK, VPREB3, CD79B, LYN and IRF8 were downregulated in both chicken and mouse (Table 5.4). The upregulated genes had less in common (Table 5.5). This finding is likely to be related to the differentiation of the DT40 and mouse pro-B cells in opposite directions upon Pax5 deletion (I; Schebesta et al. 2007; Pridans et al. 2008).

Direct comparison of individual pathway enrichment analysis results produced very few similarly regulated pathways. However, the meta-analysis (IV) was able to uncover a several putatively conserved enriched pathways from the cross-species data. This indicates that higher statistical power may be obtained by analyzing disparate experiments together than separately (Table 5.6-5.7). It can be argued that, just missing the threshold in three or more analyses is a nearly as significant a result as when each individual pathway enrichment result is significant on its own. Meta-analysis methods, such as the RankProd R/Bioconductor package (Hong et al. 2006), are thus able to utilize the increased sample size of comparative analyses more effectively and hence more of the information contained in the data than the threshold-based methods (IV).

In this study experimentally determined gene sets from the Molecular Signatures database (MSigDB) were used (Subramanian et al. 2005), including gene sets containing genes with evolutionarily conserved binding sites for miRNAs or transcription factors (Xie et al. 2005). Upregulation of the Biocarta IL7PATHWAY and ARFPATHWAY (Table 5.7) after Pax5 deletion was detected and is likely to be indicative of differentiation away from the B cell fate. Activation of pre-B-cell receptor signaling strongly induces BCL6 expression, whereas IL-7Rα -Stat5 signaling is attenuated (Malin et al. 2010a; Malin et al. 2010b). At the transition from IL-7-dependent into IL-7-independent stages of B-cell development, BCL6 is activated. The DNA breaks that occur as a result of the Ig light chain gene rearrangements lead to excessive up-regulation of Arf and p53 in the absence of BCL6 (Duy 2010). It is not clear whether Pax5 is directly involved in the repression of the IL7 pathway, but induction of the IL-7Rα was observed in the mouse as well (Delogu et al. 2006).

The motif-based gene sets indicated that the predicted targets of the miRNA Mir-503 are up-regulated upon Pax5 deletion. The actions of the Mir-503 are similar to the Mir-155 in that both regulate proliferation related targets, and their induction promotes differentiation (Forrest et al. 2010). It would be interesting to find out whether the expression of the Mir-503 is regulated by Pax5 – either directly or indirectly.

The combining of this analysis with data from ChIP-seq studies (Schmidt et al. 2010) would allow one to additionally gauge the extent to which evolutionarily conserved targets are enriched for direct target interactions (Lin et al. 2010; Treiber et al. 2010).

6.2.3. The Whole Genome Duplication and Ikaros family evolution

Because whole genome duplication (WGD) affects the organism's every gene simultaneously, it generates large amounts of genetic raw material for establishing new regulatory circuits (Van de Peer et al. 2009; Van de Peer et al. 2010; Dehal & Boore 2005). Genes retained after whole genome duplication are sometimes termed ohnologs (Nakatani et al. 2007). Studies of Ikaros family evolution (II) indicate that the Ikaros family are also ohnologs and not just paralogs of each other (II; John et al. 2009).

A recent study supports the role of the Ikaros family in the evolution of the gnastosthome adaptive immune system (John et al. 2009). This study links the evolution of the Ikaros family to the second whole genome duplication proposed by Susumo Ohno. The difference between this more recent study and Article II is that the inference of evolutionary relationships is based on local synteny, termed microsynteny, from the lamprey genome. Thus, it was possible to infer, more reliably than from the sequence alone, that the two members of the lamprey Ikaros family are likely to correspond to the ancestors of the Ikaros/Aiolos and the Helios/Eos branches, instead of having arisen from a gene duplication event that occurred after the jawed and jawless vertebrate lineages diverged (John et al. 2009). Article II uses regular or macrosynteny as well as sequence-based reconstructions to argue that the Ikaros family is likely to have doubled in size as a result of the second genome wide duplication event (2R) that took place after the divergence of the jawed and jawless vertebrates. The precise role played by the Ikaros family in the adaptive immune system evolution is not addressed by either study.

Views of blood cell formation are moving away from the requirement for discrete branch-points that would lead to a well-defined hierarchy of development (Wellner 2008; Kawamoto & Katsura 2009). A recent "myeloid based" model of hematopoietic stem cell differentiation proposes an evolutionary explanation for the observed propensity of every cell type to retain the potential for myeloid differentiation almost until the very end (Kawamoto & Katsura 2009). In the case of the B-cell lineage this is until the Pax5 expression fixes the fate of the lineage. The model divides extant cell types into specialized and prototype classes. In the myeloid-based model, the cell types such as macrophages and phagocytes are the prototype cells, and the B cells and T cells represent specialized cell types. In general, the cell types that are effector cells of the adaptive immune system are specialized, and others, except erythroid cells, tend to be prototypic.

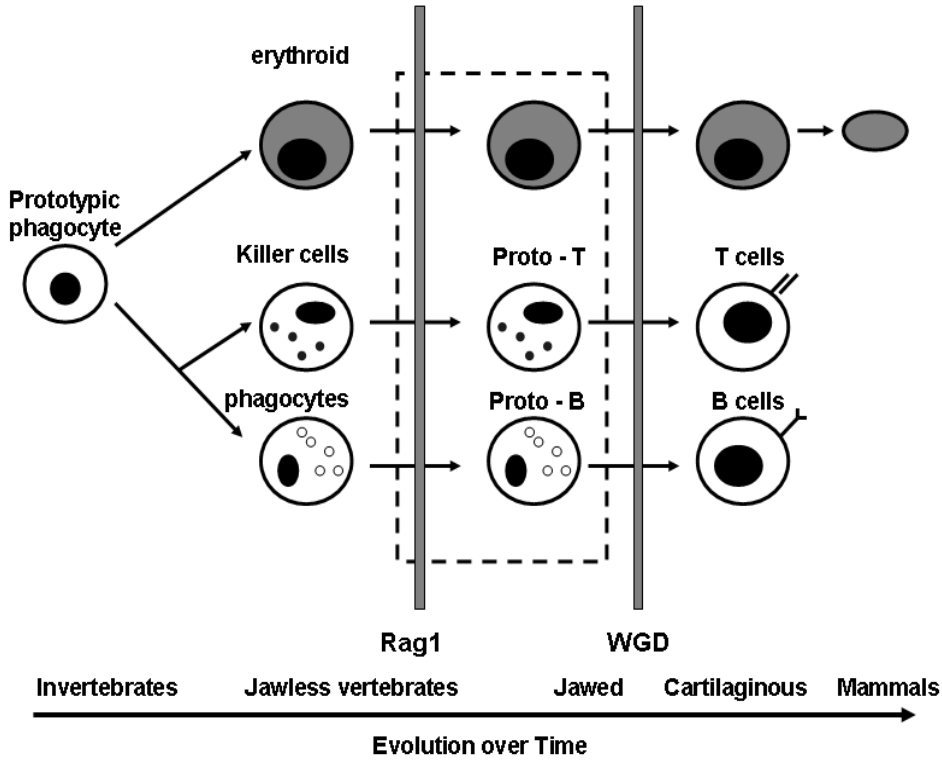

Figure 11. A model of immune system evolution that is consistent with the "myeloid model" of haematopoiesis. The acquisition of the Rag1 gene and the second Whole Genome Duplication (WGD) were probably separate events, but the genome duplication helped in the evolution of the transcriptional regulatory networks needed for the function and development of the immune system.

B cells are, according to the myeloid model, thought to have arisen from a phagocyte/macrophage-like cell (Kawamoto & Katsura 2009) (Figure 11). Accordingly, they retain the ability to present antigens and have a propensity to be converted into macrophages. T cells are hypothesized to have been derived from a cytotoxic killer cell type. The specialized cells developed when regulatory machinery was added on top of them after the Rag1 recombinase became available. The development of the regulatory machinery may have required the second genome-wide duplication event (Figure 11).

It is interesting to speculate that regulatory factors could be also divided into specialized and prototypic factors. After the WGD one factor could be thought to retain its earlier function while the other factor, if it is retained, acquires novel functions. In the Ikaros family, Ikaros is widely expressed and seems to have broader roles. Aiolos has roles in the late stages of B-cell differentiation. Since Ikaros and Aiolos, as well as Helios and Eos, are in separate branches of the phylogenetic tree, Ikaros might be seen as a prototypic member compared to Aiolos.

Some of the regulatory complexity of adaptive immunity and especially the B cell into plasma cell transition may also stem from the superimposition of more specialized

functions on top of the primitive ones, creating overlapping regulatory circuits (Lossos 2007; Van de Peer et al. 2009; Erwin & Davidson 2009; Kawamoto & Katsura 2009). This may be evident in the role of IRF4 in plasma cell differentiation (Lossos 2007; Saito et al. 2007) and in general what seem to be redundant regulatory mechanisms at this point. The plasma cell type was probably newly created, since it is so firmly anchored to adaptive immunity. The XBP-1 gene is utilized by other secretory cell types as well (Acosta-Alvear et al. 2007). So the plasma cell may have been created by the combination of immunological specialization, Rag1 acquisition, and co-opting the factors necessary for the secretory phenotype (Figure 11).

It would be interesting to see whether a completely non-immune cell, such as a fibroblast or an iPS cell, could be turned into a plasma cell by artificially recreating these conditions, namely expressing Blimp1, IFR4 and XBP1. For this to work a re-arranged immunoglobulin gene might need to be expressed as well, so that the cells would have something to secrete, which would help to induce the unfolded protein response (UPR). BCL6 also seems not to be part of the original B-cell program, but is more likely required for maintaining the high proliferation rate in the germinal centre cells in the face of DNA damage (Phan & Dalla-Favera 2004). This may be why the avian bursa of Fabricius expresses it as well (Koskela et al. 2003).

7. CONCLUDING REMARKS

The dual roles of Pax5, both at the beginning and the end of the B-cell differentiation, unfold a picture of a prototypic cell fate determining factor which both positively regulates the genes related to its own cell fate and also represses alternative cell fates – including terminal differentiation. In order to achieve terminal differentiation status such factors need to be repressed. This study shows that the inactivation of the Pax5 gene is necessary but can also be sufficient for terminal differentiation towards the plasma cell fate.

The Ikaros family of gene regulatory proteins is central throughout B-cell and T-cell development. The expansion of the Ikaros family is likely to have occurred as a result of the second genome-wide duplication. The evolution of the adaptive immune system is likely to be connected to the second genome-wide duplication as well as the acquisition of the Rag1 gene recombination system. The evolutionary path that leads to its formation could help to explain the regulatory network structure of the present day immune system.

This regulatory circuit related to the B cell program and terminal differentiation is highly conserved in evolution. The results pertaining to the functions of the circuit obtained using the DT40 B cell model systems are consistent with the human and murine results. It is likely that cross-species regulatory studies, combining ChIP-seq and gene expression with functional studies, will help to establish the logic of regulatory network functions as well as their evolution.

ACKNOWLEDGEMENTS

This work was carried out at the Turku Graduate School of Biomedical Sciences (TuBS) and at the Department of Medical Microbiology and Immunology, University of Turku, during the years 1999-2010. I am most grateful for the support, advice and encouragement over these years that made it possible for me to complete the research. Especially I wish to thank the following persons.

I want to express my heartfelt gratitude to Professor Olli Lassila for the opportunity to conduct my research in his group and for being an inspiring supervisor providing ideas, encouragement, and guidance and as well as pertinent criticism, without which my work would not have been possible. I benefited greatly from the symposia organized by Olli as the Director of TuBS, bringing world-class visiting scientists to Turku. Olli and the graduate school have also enabled me to travel to a number of international scientific conferences and attend seminars which have helped me to develop my understanding of bioinformatics, enhancing my professional education.

I thank my supervisory committee, Professor Mark Johnson and Professor Riitta Lahesmaa, for their support and inspiration in and out of supervisory committee meetings. I am also very grateful to the pre-examiners, Professor Mark Johnson and Professor Seppo Meri, for helping me to see where and how I could improve my manuscript.

I thank Professor Emeritus Matti Viljanen, Professor Emeritus Paavo Toivanen, Professor Olli Vainio, Professor Sirpa Jalkanen, Professor Emeritus Heikki Arvilommi, Professor Pentti Huovinen, Professor Marko Salmi and senior scientists Markku Viander, Erkki Eerola, Arno Hänninen, Jussi Kantele, and Juha Suhonen for their understanding discussions and perceptive questions and comments during progress reports, journal clubs and other seminars at the Department of Medical Microbiology.

Professor Riitta Lahesmaa, Professor John Eriksson, Professor Seppo Meri, Professor Olli Silvennoinen, Professor Ilkka Julkunen, Professor Jukka Pelkonen, Professor Sirpa Jalkanen, Stephen Rudd, Christophe Roos and Professor Mark Johnson provided me an innovative scientific community and were also helpful models of scientists.

I want to thank Professor Jean-Marie Buerstedde for giving us the opportunity to use the DT-40 cell line model system, which forms the basis of the main body of the work in this study. As an accomplished bioinformatician, Jean-Marie sequenced the EST libraries and established the data bases that were prerequisite for our array studies that lead to the discovery of the B cell to plasma cell transition in the Pax5 knock-out cells. I wish to extend my sincere thanks to the members of Jean-Marie's lab, including Dr. Arakawa, for their help in establishing the DT-40 technology in Turku.

I wish to thank all the members of Olli Lassila's research group for their peer-support and company: Jukka Alinikula, Jenny Granberg, Dr. Jenni Heikkinen, Janne Komi, Kimmo Koskela, Kirsi Laine, Jussi Liippo, Paulina Mikolajczak, Laura Mustonen, Milja Möttönen, Elli Narvi, Kalle-Pekka Nera, Veera Nikoskelainen, Pia Suonpää, Perttu Terho and Milja Ylihärsilä. I also thank Mari Erlin, Anna Karvonen, Mari Virta, Jasperiina Mattson, Anne Peippo and Ann Sofie Wierda for their technical assistance in the lab.

I am grateful to the TuBS staff: Nina Widberg, Ph.D. Docent Heli Salminen-Mankonen (currently at the BTK), MSc. Susanna Rosenberg, MSc. Laura Kopu and Teija Aho for their care in keeping the graduate school running. My thanks also belong to Raija Raulimo, Tuula Rikalainen, Diina Ryynänen, Matti Toivonen, Mervi Turta and Paula Vahakoski for their assistance in the secretariat of the Microbiology and Immunology Department. I also acknowledge Teuvo Virtanen's help with computers. Mika Korkeamäki and Perttu Terho have given the research project invaluable help in cytometry.

I want to thank my room-mates and brothers-in-arms over the years: Kalle-Pekka Nera as well as Kimmo Koskela and Jussi Kantele. We had fun times and mind-boggling brainstorming sessions together in the glass-ware storage room (lasivarasto). I wish to especially thank Kalle-Pekka for personal support during sometimes difficult times and credit him for establishing the DT-40 technology in the lab. I have enjoyed working with you.

I also wish to thank the Centre for Scientific Computing in Finland (CSC) for the computing resources as well as for the courses on bioinformatics, including the 2005 first-ever R/Bioconductor course in Finland. I would especially like to thank Dr. Jarno Tuimala, Dr. Kimmo Mattila and Dr. Eija Korpelainen, who provided me invaluable support and guidance as a neophyte bioinformatician.

I have worked as a bioinformatician at the VTT Medical Biotechnology Department since 2006, and would like to thank Professor Olli Kallioniemi for his inspiring leadership and Dr. Merja Perälä for encouragement and flexibility regarding my working arrangements. Harri Siitari has also given me helpful encouragement. I have been a part of a community of bioinformaticians at the VTT and have learned a great deal from them. Dr. Vidal Fey, Henrik Edgren, Tommi Pisto, Dr. Arho Virkki, John-Patrick Mpindi, Elmar Bucher and especially Henri Sara have contributed to my understanding of R programming and the Linux operating system. The biologists at the VTT, including Dr. Päivi Östling, Dr. Tao He, Dr. Saija Haapa-Paananen and Dr. Suvi-Katri Leivonen have pushed me to develop my skills in bioinformatics by providing challenging problems to solve and data to analyze. I also wish to credit Dr. Matthias Nees for inspiring me to develop the meta-analysis workflow that I used in Article IV, which will be developed further in future projects.

I also want to thank my friends in Tampere, especially Jari Antikainen, Maria Antikainen, Mikko Leino, Mikko Kallionsivu, Henrikki Määttä and Martti Laine, for the many years of friendship and shared interest in fantasy role-playing games.

My warmest thanks belong to my parents Viljo and Lilja as well as my brother Matti and sister Liisa for walking with me throughout these years and for their unfailing support.

I would like to thank the Turku Graduate School of Biomedical Sciences, the European Union (QLK3-CT-2000-00785, Apo-sys consortium, ProspeR consortium, RIGHT consortium, Genica consortium), Tekes (The National Technology Agency), the Academy of Finland, the Finnish Cultural Foundation (Kulttuurirahasto), the Turku University Foundation, Turku Microbiological Society, Emil Aaltonen Foundation, US National Science Foundation, US Department of Energy and Karjalan Sivistysseura ry for their valuable financial support during the time I have been working on my thesis.

REFERENCES

Abdrakhmanov I, Lodygin D, Geroth P, Arakawa H, Law A, Plachy J, Korn B, Buerstedde JM. A large database of chicken bursal ESTs as a resource for the analysis of vertebrate gene function. Genome Res. 2000; 10(12):2062-9.

Acosta-Alvear D, Zhou Y, Blais A, Tsikitis M, Lents NH, Arias C, Lennon CJ, Kluger Y, Dynlacht BD. XBP1 controls diverse cell type- and condition-specific transcriptional regulatory networks. Mol Cell 2007; 27(1):53-66.

Aderem A. Systems biology: its practice and challenges. Cell 2005; 121(4):511-13.

Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y. Gene prioritization through genomic data fusion. Nat Biotechnol. 2006; 24(5):537-44.

Adolfsson J, Månsson R, Buza-Vidas N, Hultquist A, Liuba K, Jensen CT, Bryder D, Yang L, Borge OJ, Thoren LA, Anderson K, Sitnicka E, Sasaki Y, Sigvardsson M, Jacobsen SE. Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential: a revised road map for adult blood lineage commitment. Cell 2005; 121(2):295-306.

Aittokallio T, Schwikowski B. Graph-based methods for analysing networks in cell biology. Brief Bioinform. 2006; 7(3):243-55.

Alder MN, Herrin BR, Sadlonova A, Stockard CR, Grizzle WE, Gartland LA, Gartland GL, Boydston JA, Turnbough CL Jr, Cooper MD. Antibody responses of variable lymphocyte receptors in the lamprey. Nat Immunol. 2008; 9(3):319-27.

Alinikula J, Lassila O, Nera KP. DT40 mutants: a model to study transcriptional regulation of B cell development and function. Subcell Biochem. 2006; 40:189-205.

Alinikula J, Kohonen P, Nera KP, Lassila O. Concerted action of Helios and Ikaros controls the expression of the inositol 5-phosphatase SHIP. Eur J Immunol. 2010; 40(9):2599-607.

Alles MC, Gardiner-Garden M, Nott DJ, Wang Y, Foekens JA, Sutherland RL, Musgrove EA, Ormandy CJ. Meta-analysis and gene set enrichment relative to ER status reveal elevated activity of MYC and E2F in the "basal" breast cancer subgroup. PLoS One. 2009; 4(3):e4710.

Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet. 2006; 7(1):55-65.

Alon U. An introduction to systems biology: Design Principles of biological circuits. London 2007: Chapman & Hall, CRC Press.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25(17):3389-402.

Amin RH, Schlissel MS. Foxo1 directly regulates the transcription of recombination-activating genes during B cell development. Nat Immunol. 2008; 9(6):613-22.

Arakawa H, Buerstedde JM. Immunoglobulin gene conversion: insights from bursal B cells and the DT40 cell line. Dev Dyn 2004; 229:458–64.

Arakawa H, Hauschild J, Buerstedde JM. Requirement of the activation-induced deaminase (AID) gene for immunoglobulin gene conversion. Science 2002; 295(5558):1301-6.

Arakawa H, Lodygin D, Buerstedde JM. Mutant loxP vectors for selectable marker recycle and conditional knock-outs. BMC Biotechnol. 2001; 1:7.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25(1):25-9.

Autio R, Kilpinen S, Saarela M, Kallioniemi O, Hautaniemi S, Astola J. Comparison of Affymetrix data normalization methods using 6,926 experiments across five array generations. BMC Bioinformatics 2009; 10 Suppl 1:S24.

Bain G, Maandag EC, Izon DJ, Amsen D, Kruisbeek AM, Weintraub BC, Krop I, Schlissel MS, Feeney AJ, van Roon M, et al. E2A proteins are required for proper B cell development and initiation of immunoglobulin gene rearrangements. Cell 1994; 79(5):885-92.

Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004; 5(2):101-13.

Barabási AL. Scale-free networks: a decade and beyond. Science 2009; 325(5939):412-13.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 2009; 37(Database issue):D885-90.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 2006; 125(2):315-26.

Blagodatski A, Batrak V, Schmidl S, Schoetz U, Caldwell RB, Arakawa H, Buerstedde JM. A cis-acting diversification activator both necessary and

sufficient for AID-mediated hypermutation. PLoS Genet. 2009; 5(1):e1000332.

Boisset JC, van Cappellen W, Andrieu-Soler C, Galjart N, Dzierzak E, Robin C. In vivo imaging of haematopoietic cells emerging from the mouse aortic endothelium. Nature 2010; 464(7285):116-20.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003; 19(2):185-93.

Boutros M, Ahringer J. The art and design of genetic screens: RNA interference. Nat Rev Genet. 2008; 9(7):554-66.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001; 29(4):365-71.

Brent MR. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat Rev Genet. 2008; 9(1):62-73.

Brown WR, Hubbard SJ, Tickle C, Wilson SA. The chicken as a model for large-scale analysis of vertebrate gene function. Nat Rev Genet. 2003; 4(2):87-98.

Bryder D, Rossi DJ, Weissman IL. Hematopoietic stem cells: the paradigmatic tissue-specific stem cell. Am J Pathol. 2006; 169(2):338-46.

Bryder D, Sigvardsson M. Shaping up a lineage-lessons from B lymphopoesis. Curr Opin Immunol. 2010; 22(2):148-53.

Buerstedde JM, Takeda S. Increased ratio of targeted to random integration after transfection of chicken B cell lines. Cell 1991; 67:179–188.

Burt DW. Chicken genome: current status and future opportunities. Genome Res. 2005; 15(12):1692-8.

Cairns BR. The logic of chromatin architecture and remodelling at promoters. Nature 2009; 461(7261):193-8.

Calame K. Activation-dependent induction of Blimp-1. Curr Opin Immunol. 2008; 20(3):259-64. Epub 2008 Jun 12.

Caldwell RB, Kierzek AM, Arakawa H, Bezzubov Y, Zaim J, Fiedler P, Kutter S, Blagodatski A, Kostovska D, Koter M, Plachy J, Carninci P, Hayashizaki Y, Buerstedde JM. Full-length cDNAs from chicken bursal lymphocytes to facilitate gene function analysis. Genome Biol. 2005; 6(1):R6.

Carlisle AJ, Prabhu VV, Elkahloun A, Hudson J, Trent JM, Linehan WM, Williams ED, Emmert-Buck MR, Liotta LA, Munson PJ, Krizman DB. Development of a prostate cDNA microarray and statistical gene expression analysis package. Mol. Carcinog. 2000; 28:12–22.

Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe MJ, Wilde A, Brudno M, Morris QD, Hughes TR. Conservation of core gene expression in vertebrate tissues. J Biol. 2009; 8(3):33.

Chen Y, Dougherty E, Bittner M. Ratio-based decisions and the quantitative analysis of cDNA microarray images. J Biomed Opt.1997; 2:364–374.

Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer 2008; 8(1):37-49.

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD. Integration of biological networks and gene expression data using Cytoscape. Nat Protoc. 2007; 2(10):2366-82.

Cobaleda C, Jochum W, Busslinger M. Conversion of mature B cells into T cells by dedifferentiation to uncommitted progenitors. Nature 2007a; 449(7161):473-7.

Cobaleda C, Schebesta A, Delogu A, Busslinger M. Pax5: the guardian of B cell identity and function. Nat Immunol. 2007b; 8(5):463-70.

Cooper MD, Alder MN. The evolution of adaptive immune systems. Cell 2006; 124(4):815-22.

Cooper MD, Peterson RD, Good RA. Delineation of the thymic and bursal lymphoid systems in the chicken. Nature 1965; 205:143-6.

Darwin C (1872). The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life (6th ed.). London: John Murray. ISBN 1435393864. Retrieved 2009-11-01.

De Vos J, Hose D, Rème T, Tarte K, Moreaux J, Mahtouk K, Jourdan M, Goldschmidt H, Rossi JF, Cremer FW, Klein B. Microarray-based understanding of normal and malignant plasma cells. Immunol Rev. 2006; 210:86-104.

Decker T, Pasca di Magliano M, McManus S, Sun Q, Bonifer C, Tagoh H, Busslinger M. Stepwise activation of enhancer and promoter regions of the B cell commitment gene Pax5 in early lymphopoiesis. Immunity 2009; 30(4):508-20.

Dehal P, Boore JL. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 2005; 3(10):e314.

DeKoter RP, Singh H. Regulation of B lymphocyte and macrophage development by graded expression of PU.1 Science 2000; 288(5470):1439-41.

Delogu A, Schebesta A, Sun Q, Aschenbrenner K, Perlot T, Busslinger M. Gene repression by Pax5 in B cells is essential for blood cell homeostasis and

is reversed in plasma cells. Immunity 2006; 24(3):269-81.

Dengler HS, Baracho GV, Omori SA, Bruckner S, Arden KC, Castrillon DH, DePinho RA, Rickert RC. Distinct functions for the transcription factor Foxo1 at various stages of B-cell differentation. Nat Immunol. 2008; 9(12):1388-98.

Dias S, Månsson R, Gurbuxani S, Sigvardsson M, Kee, BL. E2A proteins promote development of lymphoid-primed multipotent progenitors. Immunity 2008; 29(2):217-27.

Draghici S, Khatri P, Eklund AC, Szallasi Z. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet. 2006; 22(2):101-9.

Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO, Alizadeh AA. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Res. 2003; 31(1):219-23.

Durand C, Dzierzak E. Embryonic beginnings of adult hematopoietic stem cells. Haematologica 2005; 90: 100-108.

Durbin B, Rocke DM. Estimation of transformation parameters for microarray data. Bioinformatics 2003; 19(11):1360-7.

Duy C, Yu JJ, Nahar R, Swaminathan S, Kweon SM, Polo JM, Valls E, Klemm L, Shojaee S, Cerchietti L, Schuh W, Jäck HM, Hurtz C, Ramezani-Rad P, Herzog S, Jumaa H, Koeffler HP, de Alborán IM, Melnick AM, Ye BH, Müschen M. BCL6 is critical for the development of a diverse primary B cell repertoire. J Exp Med. 2010; 207(6):1209-21.

Dysvik B, Jonassen I. J-Express: exploring gene expression data using Java. Bioinformatics 2001; 17(4):369-70.

Eberhard D, Jiménez G, Heavey B, Busslinger M. Transcriptional repression by Pax5 (BSAP) through interaction with corepressors of the Groucho family. EMBO J. 2000; 19(10):2292-303.

Edwards D. Non-linear normalization and background correction in one-channel cDNA microarray studies. Bioinformatics 2003; 19(7):825-33.

Efron B. The Jackknife, the Bootstrap and Other Resampling Plans. CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38, SIAM, Philadelphia 1982.

Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. PNAS 1996; 93(23): 13429-34.

Efron B, Tibshirani, R. On testing the significance of sets of genes. Stanford tech report rep 2006. http://www-stat.stanford.edu/~tibs/ftp/GSA.pdf

Eilken HM, Nishikawa S, Schroeder T. Continuous single-cell imaging of blood generation from haemogenic endothelium. Nature 2009; 457(7231):896-900.

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998; 95(25):14863-8.

Elo LL, Järvenpää H, Oresic M, Lahesmaa R, Aittokallio T. Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. Bioinformatics 2007; 23(16):2096-103. Epub 2007 Jun 6.

Erfle H, Neumann B, Liebel U, Rogers P, Held M, Walter T, Ellenberg J, Pepperkok R. Reverse transfection on cell arrays for high content screening microscopy. Nat Protoc. 2007; 2(2):392-9.

Erwin DH, Davidson EH. The evolution of hierarchical gene regulatory networks. Nat Rev Genet. 2009; 10(2):141-8.

Ettwiller L, Budd A, Spitz F, Wittbrodt J. Analysis of mammalian gene batteries reveals both stable ancestral cores and highly dynamic regulatory sequences. Genome Biol. 2008; 9(12):R172.

Fairfax KA, Kallies A, Nutt SL, Tarlinton DM. Plasma cell development: from B-cell subsets to long-term survival niches. Semin Immunol. 2008; 20(1):49-58.

Farnham PJ. Insights from genomic profiling of transcription factors. Nat Rev Genet. 2009; 10(9):605-16.

Felsenstein J. PHYLIP (Phylogeny Inference Package), Version3.6b [Computer Program]. Distributed by the author. Seattle 2004: Department of Genome Sciences, University of Washington.

Feret J, Danos V, Krivine J, Harmer R, Fontana W. Internal coarse-graining of molecular systems. Proc Natl Acad Sci U S A. 2009; 106(16):6453-8.

Fillatreau S, Radbruch A. IRF4 - a factor for class switching and antibody secretion. Nat Immunol. 2006; 7(7):704-6.

Forrest AR, Kanamori-Katayama M, Tomaru Y, Lassmann T, Ninomiya N, Takahashi Y, de Hoon MJ, Kubosaki A, Kaiho A, Suzuki M, Yasuda J, Kawai J, Hayashizaki Y, Hume DA, Suzuki H. Induction of microRNAs, mir-155, mir-222, mir-424 and mir-503, promotes monocytic differentiation through combinatorial regulation. Leukemia 2010; 24(2):460-6.

Fujita N, Jaye DL, Geigerman C, Akyildiz A, Mooney MR, Boss JM, Wade PA. MTA3 and the Mi-2/NuRD complex regulate cell fate during B lymphocyte differentiation. Cell 2004; 119(1):75-86.

Furlong RF. Insights into vertebrate evolution from the chicken genome sequence. Genome Biol. 2005; 6(2):207.

Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy: Analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 2004; 20(3):307-15.

Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res. 2010; 38(Database issue):D331-5.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004; 5(10):R80.

Georgopoulos K. Haematopoietic cell-fate decisions, chromatin regulation and ikaros. Nat Rev Immunol. 2002; 2(3):162-74.

Georgopoulos K, Bigby M, Wang JH, Molnar A, Wu P, Winandy S, Sharpe A. The Ikaros gene is required for the development of all lymphoid lineages. Cell 1994; 79(1):143-56.

Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. Widespread monoallelic expression on human autosomes. Science 2007; 318(5853):1136-40.

Graur D, Li WH. Fundamentals of molecular evolution. Massachusetts 2000: Sinauer Associates.

Grunstein M, Hogness DS. Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. Proc Natl Acad Sci U S A. 1975; 72(10):3961-5.

Halperin Y, Linhart C, Ulitsky I, Shamir R. Allegro: analyzing expression and sequence in concert to discover regulatory programs. Nucleic Acids Res. 2009; 37(5):1566-79.

Hardy RR, Kincade PW, Dorshkind K. The protean nature of cells in the B lymphocyte lineage. Immunity 2007; 26(6):703-14.

Hauser J, Verma-Gaur J, Wallenius A, Grundström T. Initiation of antigen receptor-dependent differentiation into plasma cells by calmodulin inhibition of E2A. J Immunol. 2009; 183(2):1179-87.

Henney A, Superti-Furga G. A network solution. Nature 2008; 455(7214):730-1.

Hoffmann R, Lottaz C, Kühne T, Rolink A, Melchers F. Neutrality, compensation, and negative selection during evolution of B-cell development transcriptomes. Mol Biol Evol. 2007; 24(12):2610-8.

Hoffmann R, Melchers F. A genomic view of lymphocyte development. Curr Opin Immunol. 2003; 15(3):239-45.

Hoffmann R, Seidl T, Neeb M, Rolink A, Melchers F. Changes in gene expression profiles in developing B cells of murine bone marrow. Genome Res. 2002; 12(1):98-111.

Hoheisel JD. Microarray technology: beyond transcript profiling and genotype analysis. Nat Rev Genet. 2006; 7(3):200-10.

Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics 2006; 22(22):2825-7.

Hood L, Heath JR, Phelps ME, Lin B. Systems biology and new technologies enable predictive and preventative medicine. Science 2004; 306(5696):640-3.

Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol. 2008; 4(11):682-90.

Hosack DA, Dennis G Jr, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biol. 2003; 4(10):R70.

Houssaint E, Lassila O, Vainio O. Bu-1 antigen expression as a marker for B cell precursors in chicken embryos. Eur J Immunol. 1989; 19:239–243.

Hu M, Krause D, Greaves M, Sharkis S, Dexter M, Heyworth C, Enver T. Multilineage gene expression precedes commitment in the hemopoietic system. Genes Dev. 1997; 11(6):774-85.

Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4(1):44-57.

Hubbard SJ, Grafham DV, Beattie KJ, Overton IM, McLaren SR, Croning MD, Boardman PE, Bonfield JK, Burnside J, Davies RM, Farrell ER, Francis MD, Griffiths-Jones S, Humphray SJ, Hyland C, Scott CE, Tang H, Taylor RG, Tickle C, Brown WR, Birney E, Rogers J, Wilson SA. Transcriptome analysis for the chicken based on 19,626 finished cDNA sequences and 485,337 expressed sequence tags. Genome Res. 2005; 15(1):174-83.

Huber PJ, Robust Statistics. New York 1981: Wiley.

Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 2002; 18 Suppl 1:S96-104.

Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet. 2001; 2:343-72.

International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 2004; 432(7018):695-716.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003; 4(2):249-64.

Ivanova NB, Dimos JT, Schaniel C, Hackney JA, Moore KA, Lemischka IR. A stem cell molecular signature. Science 2002; 298(5593):601-4.

Janeway C, Murphy KP, Travers P, Walport M. Janeway's immuno biology. New York 2008: Garland Science.

John LB, Yoong S, Ward AC. Evolution of the Ikaros gene family: implications for the origins of adaptive immunity. J Immunol. 2009; 182(8):4792-9.

Joyce AR, Palsson BØ. The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol. 2006; 7(3):198-210.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005; 110:462-467.

Kallies A, Hasbold J, Fairfax K, Pridans C, Emslie D, McKenzie BS, Lew AM, Corcoran LM, Hodgkin PD, Tarlinton DM, Nutt SL. Initiation of plasma-cell differentiation is independent of the transcription factor Blimp-1. Immunity 2007; 26(5):555-66.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2008; 36(Database issue):D480-4.

Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28(1):27-30.

Kapur K, Xing Y, Ouyang Z, Wong WH. Exon arrays provide accurate assessments of gene expression. Genome Biol. 2007; 8(5):R82.

Kapushesky M, Kemmeren P, Culhane AC, Durinck S, Ihmels J, Körner C, Kull M, Torrente A, Sarkans U, Vilo J, Brazma A. Expression Profiler: next generation - an online platform for analysis of microarray data. Nucleic Acids Res. 2004; 32(Web Server issue):W465-70.

Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol. 2008; 9(10):770-80.

Kasahara M. The 2R hypothesis: an update. Curr Opin Immunol. 2007; 19(5):547-52.

Kaufman L, Rousseeuw PJ. Finding groups in data: An introduction to cluster analysis. New York 1990: Wiley.

Kawamoto H, Katsura Y. A new paradigm for hematopoietic cell lineages: revision of the classical concept of the myeloid-lymphoid dichotomy. Trends Immunol. 2009; 30(5):193-200.

Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. J Comput Biol. 2000; 7(6):819-37.

Kilpinen S, Autio R, Ojala K, Iljin K, Bucher E, Sara H, Pisto T, Saarela M, Skotheim RI, Björkman M, Mpindi JP, Haapa-Paananen S, Vainio P, Edgren H, Wolf M, Astola J, Nees M, Hautaniemi S, Kallioniemi O. Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. Genome Biol. 2008; 9(9):R139. Epub 2008 Sep 19.

Kim HD, Shay T, O'Shea EK, Regev A. Transcriptional Regulatory Circuits: Predicting Numbers from Alphabets. Science 2009; 325(5939):429-32.

Kirschner MW. The meaning of systems biology. Cell 2005; 121(4):503-4.

Klein U, Casola S, Cattoretti G, Shen Q, Lia M, Mo T, Ludwig T, Rajewsky K, Dalla-Favera R. Transcription factor IRF4 controls plasma cell differentiation and class-switch recombination. Nat Immunol. 2006; 7(7):773-82.

Klein U, Dalla-Favera R. Germinal centres: role in B-cell physiology and malignancy. Nat Rev Immunol. 2008; 8(1):22-33.

Klein U, Dalla-Favera R. Unexpected steps in plasma-cell differentiation. Immunity 2007; 26(5):543-4.

Komili S, Silver PA. Coupling and coordination in gene expression processes: a systems biology view. Nat Rev Genet. 2008; 9(1):38-48.

Koskela K, Arstila T, Lassila O. Costimulatory function of CD28 in avian γδ T cells is evolutionarily conserved. Scand J Immunol. 1998; 48:635-41.

Koskela K, Kohonen P, Nieminen P, Buerstedde JM, Lassila O. Insight into lymphoid development by gene expression profiling of avian B cells. Immunogenetics 2003; 55:412-22.

Koskela K, Kohonen P, Salminen H, Uchida T, Buerstedde JM, Lassila O. Identification of a novel cytokine-like transcript differentially expressed in avian gammadelta T cells. Immunogenetics 2004; 55(12):845-54.

Kramer R, Cohen D. Functional genomics to new drug targets. Nat Rev Drug Discov. 2004; 3(11):965-72.

Kreil DP, Russell RR. There is no silver bullet - a guide to low-level data transforms and normalisation methods for microarray data. Brief Bioinform. 2005; 6(1):86-97.

Kurosaki T. Regulation of B-cell signal transduction by adaptor proteins. Nat Rev Immunol. 2002; 2:354-63.

Kwon H, Thierry-Mieg D, Thierry-Mieg J, Kim HP, Oh J, Tunyaplin C, Carotta S, Donovan CE, Goldman ML, Tailor P, Ozato K, Levy DE, Nutt SL, Calame K, Leonard WJ. Analysis of interleukin-21-induced Prdm1 gene regulation reveals functional cooperation of STAT3 and IRF4 transcription factors. Immunity 2009; 31(6):941-52.

Kwon K, Hutter C, Sun Q, Bilic I, Cobaleda C, Malin S, Busslinger M. Instructive role of the transcription factor E2A in early B lymphopoiesis and germinal center B cell development. Immunity 2008; 28(6):751-62.

Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A,

Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006; 313(5795):1929-35.

Lamb J. The Connectivity Map: a new tool for biomedical research. Nat Rev Cancer 2007; 7(1):54-60.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. Initial sequencing and analysis of the human genome. Nature 2001; 409(6822):860-921.

Laslo P, Spooner CJ, Warmflash A, Lancki DW, Lee HJ, Sciammas R, Gantner BN, Dinner AR, Singh H. Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. Cell 2006; 126(4):755-66.

Laslo P, Pongubala JM, Lancki DW, Singh H. Gene regulatory networks directing myeloid and lymphoid cell fates within the immune system. Semin Immunol. 2008; 20(4):228-35.

Lassila O, Eskola J, Toivanen P, Martin C, Dieterlen-Lievre F. The origin of lymphoid stem cells studied in chick yold sac-embryo chimaeras. Nature 1978; 272(5651):353-4.

Leivonen SK, Mäkelä R, Ostling P, Kohonen P, Haapa-Paananen S, Kleivi K, Enerly E, Aakula A, Hellström K, Sahlberg N, Kristensen VN, Børresen-Dale AL, Saviranta P, Perälä M, Kallioniemi O. Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines. Oncogene 2009; 28(44):3926-36.

Liippo J, Lassila O. Avian Ikaros gene is expressed early in embryogenesis. Eur J Immunol. 1997; 27(8):1853-7.

Liippo J, Mansikka A, Lassila O. The evolutionarily conserved avian Aiolos gene encodes alternative isoforms. Eur J Immunol. 1999; 29: 2651–7.

Lin H, Grosschedl R. Failure of B-cell differentiation in mice lacking the transcription factor EBF. Nature 1995; 376(6537):263-7.

Lin KI, Angelin-Duclos C, Kuo TC, Calame K. Blimp-1-dependent repression of Pax-5 is required for differentiation of B cells to immunoglobulin M-secreting plasma cells. Mol Cell Biol. 2002; 22(13):4771-80.

Lin SM, Du P, Huber W, Kibbe WA. Model-based variance-stabilizing transformation for Illumina microarray data. Nucleic Acids Res. 2008; 36(2):e11.

Lin YC, Jhunjhunwala S, Benner C, Heinz S, Welinder E, Mansson R, Sigvardsson M, Hagman J, Espinoza CA, Dutkowski J, Ideker T, Glass CK, Murre C. A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. Nat Immunol. 2010; 11(7):635-43.

Linderson Y, Eberhard D, Malin S, Johansson A, Busslinger M, Pettersson S. Corecruitment of the Grg4 repressor by PU.1 is critical for Pax5-mediated repression of B-cell-specific genes. EMBO Rep. 2004; 5(3):291-6.

Liu ET. Integrative biology - a strategy for systems biomedicine. Nat Rev Genet. 2009; 10(1):64-8.

Liu ET. Systems biology, integrative biology, predictive biology. Cell 2005; 121(4):505-6.

Liu Q, Dinu I, Adewale AJ, Potter JD, Yasui Y. Comparative evaluation of gene-set analysis methods. BMC Bioinformatics 2007; 8:431.

Lossos IS. The endless complexity of lymphocyte differentiation and lymphomagenesis: IRF-4 downregulates BCL6 expression. Cancer Cell 2007; 12(3):189-91.

Lu R. Interferon regulatory factor 4 and 8 in B-cell development. Trends Immunol. 2008; 29(10):487-92.

Malin S, McManus S, Busslinger M. STAT5 in B cell development and leukemia. Curr Opin Immunol. 2010a; 22(2):168-76.

Malin S, McManus S, Cobaleda C, Novatchkova M, Delogu A, Bouillet P, Strasser A, Busslinger M. Role of STAT5 in controlling cell survival and immunoglobulin gene recombination during pro-B cell development. Nat Immunol. 2010b; 11(2):171-9.

Manoli T, Gretz N, Gröne HJ, Kenzelmann M, Eils R, Brors B. Group testing for pathway analysis improves comparability of different microarray datasets. Bioinformatics 2006; 22(20):2500-6.

Mansikka A, Sandberg M, Veromaa T, Vainio O, Granfors K, Toivanen P. B cell maturation in the chicken Harderian gland. J Immunol. 1989; 142:1826–1833.

Mansikka A, Sandberg M, Lassila O, Toivanen P. Rearrangement of immunoglobulin light chain genes in the chicken occurs prior to colonization of the embryonic bursa of Fabricius. Proc Natl Acad Sci U S A. 1990; 87(23):9416-20.

Martins G, Calame K. Regulation and functions of Blimp-1 in T and B lymphocytes. Annu Rev Immunol. 2008; 26:133-69.

Mayr E. What makes Biology unique? Considerations on the autonomy of a scientific discipline. Cambridge 2004: Cambridge University Press.

Medina M. Genomes, phylogeny, and evolutionary systems biology. Proc Natl Acad Sci U S A. 2005; 102 Suppl 1:6630-5.

Mikkola I, Heavey B, Horcher M, Busslinger M. Reversion of B cell commitment upon loss of Pax5 expression. Science 2002; 297(5578):110-3.

Miron M, Nadon R. Inferential literacy for experimental high-throughput biology. Trends Genet. 2006; 22(2):84-9.

Miyamoto T, Iwasaki H, Reizis B, Ye M, Graf T, Weissman IL, Akashi K. Myeloid or lymphoid promiscuity as a critical step in hematopoietic lineage commitment. Dev Cell 2002; 3(1):137-47.

Monod J, Jacob F. Cold Spring Harb. Symp. Quant. Biol. 1961; 26, 389–401.

Mora-López F, Reales E, Brieva JA, Campos-Caro A. Human BSAP and BLIMP1 conform an autoregulatory feedback loop. Blood 2007; 110(9):3150-7.

Muller T, Vingron M. Modeling amino acid replacement. J Comput Biol. 2000; 7:761–76.

Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, Dalton JD, Girtman K, Mathew S, Ma J, Pounds SB, Su X, Pui CH, Relling MV, Evans WE, Shurtleff SA, Downing JR. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. Nature 2007; 446(7137):758-64.

Murie C, Woody O, Lee AY, Nadon R. Comparison of small n statistical tests of differential expression applied to microarrays. BMC Bioinformatics 2009; 3;10(1):45.

Nakatani Y, Takeda H, Kohara Y, Morishita S. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. Genome Res. 2007; 17(9):1254-65.

Nam D, Kim SY. Gene-set approach for expression pattern analysis. Brief Bioinform. 2008; 9(3):189-97.

Neiman PE, Gehly EB, Carlson LM, Cotter RC, Thompson CB. Bursal stem cells as targets for myc-induced preneoplastic proliferation and maturation arrest. Curr Top Microbiol Immunol. 1988; 141:67–74.

Nera K-P, Alinikula J. Terho P, Narvi E, Törnquist K, Kurosaki T, Buerstedde J-M, Lassila O. Ikaros has a crucial role in regulation of B-cell receptor signaling. Eur. J. Immunol. 2006; 36:516–525.

Ng SY, Yoshida T, Zhang J, Georgopoulos K. Genome-wide lineage-specific transcriptional networks underscore Ikaros-dependent lymphoid priming in hematopoietic stem cells. Immunity 2009; 30(4):493-507.

Nieminen P, Liippo J, Lassila O. Pax-5 and EBF are expressed in committed B-cell progenitors prior to the colonization of the embryonic bursa of fabricius. Scand J Immunol. 2000; 52:465–469.

Notredame C, Higgins DG, Heringa J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000; 302:205–17.

Nutt SL, Urbánek P, Rolink A, Busslinger M. Essential functions of Pax5 (BSAP) in pro-B cell development: difference between fetal and adult B lymphopoiesis and reduced V-to-DJ recombination at the IgH locus. Genes Dev. 1997; 11(4):476-91.

Nutt SL, Morrison AM, Dörfler P, Rolink A, Busslinger M. Identification of BSAP (Pax-5) target genes in early B-cell development by loss- and gain-of-function experiments. EMBO J. 1998; 17(8):2319-33.

Nutt SL, Busslinger M. Monoallelic expression of Pax5: a paradigm for the haploinsufficiency of mammalian Pax genes? Biol Chem. 1999; 380(6):601-11.

Nutt SL, Heavey B, Rolink AG, Busslinger M. Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. Nature 1999b; 401(6753):556-62.

Ohno S. Evolution by Gene Duplication. New York 1970: Springer-Verlag.

Ollila J, Vihinen M. Immunological systems biology: gene expression analysis of B-cell development in Ramos B-cells. Mol Immunol. 2007; 44(14):3537-51.

Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics 2002; 1(5):376-86.

O'Riordan M, Grosschedl R. Coordinate regulation of B-cell differentation by the transcription factors EBF and E2A. Immunity 1999; 11(1):21-31.

Orkin SH. Diversification of haematopoietic stem cells to specific lineages. Nat Rev Genet. 2000; 1(1):57-64.

Page RD. TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci. 1996; 12:357-8.

Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009; 10(10):669-80.

Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone SA, Sklyar N, Zhao M, Sarkans U, Brazma A. ArrayExpress update - from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res. 2009; 37(Database issue):D868-72.

Parsons JD, Rodriguez-Tomé P. JESAM: CORBA software components to create and publish EST alignments and clusters. Bioinformatics 2000; 16(4):313-25.

Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci U S A. 1994; 91(11):5022-6.

Perdomo J, Holmes M, Chong B, Crossley M. Eos and pegasus, two members of the Ikaros family of proteins with distinct DNA binding activities. J Biol Chem. 2000; 275:38347-54.

Phan RT, Dalla-Favera R. The BCL6 proto-oncogene suppresses p53 expression in germinal-centre B cells. Nature 2004; 432(7017):635-9.

Pontius JU, Wagner L, Schuler GD. UniGene: a unified view of the transcriptome. The NCBI Handbook. Bethesda MD 2003: National Center for Biotechnology Information.

Pridans C, Holmes ML, Polli M, Wettenhall JM, Dakic A, Corcoran LM, Smyth GK, Nutt SL. Identification of Pax5 target genes in early B-cell differentation. J Immunol. 2008; 180(3):1719-28.

R Development Core Team. R: A language and environment for statistical computing. Vienna 2009: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ramialison M, Bajoghli B, Aghaallaei N, Ettwiller L, Gaudan S, Wittbrodt B, Czerny T, Wittbrodt J. Rapid identification of PAX2/5/8 direct downstream targets in the otic vesicle by combinatorial use of bioinformatics tools. Genome Biol. 2008; 9(10):R145.

Ramírez J, Lukin K, Hagman J. From hematopoietic progenitors to B cells: mechanisms of lineage restriction and commitment. Curr Opin Immunol. 2010; 22(2):177-84.

Rasche A, Al-Hasani H, Herwig R. Meta-analysis approach identifies candidate genes and associated molecular networks for type-2 diabetes mellitus. BMC Genomics 2008; 9:310.

Rebollo A, Schmitt C. Ikaros, Aiolos and Helios: transcription regulators and lymphoid malignancies. Immunol Cell Biol. 2003; 81(3):171-5.

Reed JL, Famili I, Thiele I, Palsson BO. Towards multidimensional genome annotation. Nat Rev Genet. 2006; 7(2):130-41.

Rhee SY, Wood V, Dolinski K, Draghici S., Use and misuse of the gene ontology annotations. Nat Rev Genet. 2008; 9(7):509-15.

Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincead-Beal C, Kulkarni P, Varambally S, Ghosh D, Chinnaiyan AM. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia 2007; 9(2):166-80.

Roessler S, György I, Imhof S, Spivakov M, Williams RR, Busslinger M, Fisher AG, Grosschedl R. Distinct promoters mediate the regulation of Ebf1 gene expression by interleukin-7 and Pax5. Mol Cell Biol. 2007; 27(2):579-94.

Rolink AG, Nutt SL, Melchers F, Busslinger M. Long-term in vivo reconstitution of T-cell development by Pax5-deficient B-cell progenitors. Nature 1999; 401(6753):603-6.

Romanow WJ, Langerak AW, Goebel P, Wolvers-Tettero IL, van Dongen JJ, Feeney AJ, Murre C. E2A and EBF act in synergy with the V(D)J recombinase to generate a diverse immunoglobulin repertoire in nonlymphoid cells. Mol Cell 2000; 5(2):343-53.

Rothenberg E. Cell lineage regulators in B and T cell development. Nat Immunol. 2007; 8(5):441-444.

Rothenberg E. B-cell specification from the genome up. Nat Immunol. 2010; 11(7):572-574.

Saito M, Gao J, Basso K, Kitagawa Y, Smith PM, Bhagat G, Pernis A, Pasqualucci L, Dalla-Favera R. A signaling pathway mediating downregulation of BCL6 in germinal center B cells is blocked by BCL6 gene alterations in B cell lymphoma. Cancer Cell 2007; 12(3):280-92.

Schadt EE. Molecular networks as sensors and drivers of common human diseases. Nature 2009; 461(7261):218-23.

Schebesta A, McManus S, Salvagiotto G, Delogu A, Busslinger GA, Busslinger M. Transcription factor Pax5 activates the chromatin of key genes involved in B cell signaling, adhesion, migration, and immune function. Immunity 2007; 27(1):49-63.

Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995; 270(5235):467-70.

Schliephake DE, Schimpl A. Blimp-1 overcomes the block in IgM secretion in lipopolysaccharide/anti-mu F(ab')2-co-stimulated B lymphocytes. Eur J Immunol. 1996; 26(1):268-71.

Schmidlin H, Diehl SA, Blom B. New insights into the regulation of human B-cell differentiation. Trends Immunol. 2009; 30(6):277-85.

Schmidlin H, Diehl SA, Nagasawa M, Scheeren FA, Schotte R, Uittenbogaart CH, Spits H, Blom B. Spi-B inhibits human plasma cell differentiation by repressing BLIMP1 and XBP-1 expression. Blood 2008; 112(5):1804-12.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. TREEPUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 2002; 18:502–4.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, Talianidis I, Flicek P, Odom DT. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 2010; 328(5981):1036-40.

Schneider P, MacKay F, Steiner V, Hofmann K, Bodmer JL, Holler N, Ambrose C, Lawton P, Bixler S, Acha-Orbea H, Valmori D, Romero P, Werner-Favre C, Zubler RH, Browning JL, Tschopp J. BAFF, a novel ligand of the tumor necrosis factor family, stimulates B cell growth. J Exp Med. 1999; 189(11):1747-56.

Scott EW, Simon J, Anastasi MC, Singh H. Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. Science 1994; 265:1573–1577.

Scott, EW, Fisher RC, Olson MC, Kehrli EW, Simon MC, Singh H. PU.1 functions in a cell-autonomous manner to control the differentiation of multipotential lymphoid-myeloid progenitors. Immunity 1997; 6:437–447.

Shaffer AL, Lin KI, Kuo TC, Yu X, Hurt EM, Rosenwald A, Giltnane JM, Yang L, Zhao H, Calame K, Staudt LM. Blimp-1 orchestrates plasma cell differentiation by extinguishing the

mature B cell gene expression program. Immunity 2002; 17(1):51-62.

Shaffer AL, Rosenwald A, Hurt EM, Giltnane JM, Lam LT, Pickeral OK, Staudt LM. Signatures of the immune response. Immunity 2001; 15(3):375-85.

Shaffer AL, Shapiro-Shelef M, Iwakoshi NN, Lee AH, Qian SB, Zhao H, Yu X, Yang L, Tan BK, Rosenwald A, Hurt EM, Petroulakis E, Sonenberg N, Yewdell JW, Calame K, Glimcher LH, Staudt LM. XBP1, downstream of Blimp-1, expands the secretory apparatus and other organelles, and increases protein synthesis in plasma cell differentiation. Immunity 2004; 21(1):81-93.

Shaffer AL, Wright G, Yang L, Powell J, Ngo V, Lamy L, Lam LT, Davis RE, Staudt LM. A library of gene expression signatures to illuminate normal and pathological lymphoid biology. Immunol Rev. 2006; 210:67-85.

Shaffer JP. Multiple hypothesis testing. Annu Rev Psychol. 1995; 46:561-84.

Shapiro-Shelef M, Calame K. Regulation of plasma-cell development. Nat Rev Immunol. 2005; 5(3):230-42.

Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008; 26(10):1135-45.

Shi L, Jones WD, Jensen RV, Harris SC, Perkins RG, Goodsaid FM, Guo L, Croner LJ, Boysen C, Fang H, Qian F, Amur S, Bao W, Barbacioru CC, Bertholet V, Cao XM, Chu TM, Collins PJ, Fan XH, Frueh FW, Fuscoe JC, Guo X, Han J, Herman D, Hong H, Kawasaki ES, Li QZ, Luo Y, Ma Y, Mei N, Peterson RL, Puri RK, Shippy R, Su Z, Sun YA, Sun H, Thorn B, Turpaz Y, Wang C, Wang SJ, Warrington JA, Willey JC, Wu J, Xie Q, Zhang L, Zhang L, Zhong S, Wolfinger RD, Tong W. The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. BMC Bioinformatics 2008; 9 Suppl 9:S10.

Smith J, Speed D, Law AS, Glass EJ, Burt DW. In-silico identification of chicken immune-related genes. Immunogenetics 2004; 56(2):122-33.

Smithyman AM, Carr K, Forman D, White RG. Separation of germinal centres from chicken spleen. Adv Exp Med Biol. 1979; 114:37–41.

Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004; 3:Article 3.

Smyth GK. Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), New York 2005: Springer.

Southern EM. DNA chips: analysing sequence by hybridization to oligonucleotides on a large scale. Trends Genet. 1996; 12(3):110-5.

Staudt LM. Cancer: negative feedback for B cells. Nature 2004; 431(7011):919-20.

Storey JD, Tibshirani R., Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003; 100(16):9440-5.

Storey JD. The positive false discovery rate: A Bayesian interpretation and the q-value. The Annals of Statistics 2003; Vol. 31, No. 6, 2013–2035.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102(43):15545-50.

Sun Y, Li H, Liu Y, Shin S, Mattson MP, Rao MS, Zhan M. Cross-species transcriptional profiles establish a functional portrait of embryonic stem cells. Genomics 2007; 89:22–35.

Suonpää P, Kohonen P, Koskela K, Koskiniemi H, Salminen-Mankonen H, Lassila O. Development of early PCLP1-expressing haematopoietic cells within the avian dorsal aorta. Scand J Immunol. 2005; 62(3):218-23.

Sweet-Cordero A, Mukherjee S, Subramanian A, You H, Roix JJ, Ladd-Acosta C, Mesirov J, Golub TR, Jacks T. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nat Genet. 2005; 37(1):48-55.

Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994; 22(22):4673-80.

Todd DJ, Lee AH, Glimcher LH. The endoplasmic reticulum stress response in immunity and autoimmunity. Nat Rev Immunol. 2008; 8(9):663-74.

Toivanen P, Toivanen A. Bursal and postbursal stem cells in chicken. Functional characteristics. Eur J Immunol. 1973; 3(9):585-95.

Treiber T, Mandel EM, Pott S, Györy I, Firner S, Liu ET, Grosschedl R. Early B Cell Factor 1 Regulates B Cell Gene Networks by Activation, Repression, and Transcription-Independent Poising of Chromatin. Immunity 2010; 32(5):714-25.

Troyanskaya OG. Putting microarrays in a context: integrated analysis of diverse biological data. Brief Bioinform. 2005; 6(1):34-43.

Turner CA Jr, Mack DH, Davis MM. Blimp-1, a novel zinc finger-containing protein that can drive the maturation of B lymphocytes into immunoglobulin-secreting cells. Cell 1994; 77(2):297-306.

Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001; 98(9):5116-21.

Van de Peer Y, Maere S, Meyer A. 2R or not 2R is not the question anymore. Nat Rev Genet. 2010; 11(2):166.

Van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 2009; 10(10):725-32.

Veistinen E, Lassila O. Bursa of Fabricius. In: Encyclopedia of Life Sciences. Chichester 2005: John Wiley; http://www.els.net: doi: 10.1038/npg.els.0003974.

Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth edition. New York 2002: Springer.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001; 291(5507):1304-51.

Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature 2009; 461(7261):199-205.

Wahl MB, Caldwell RB, Kierzek AM, Arakawa H, Eyras E, Hubner N, Jung C, Soeldenwagner M, Cervelli M, Wang YD, Liebscher V, Buerstedde JM. Evaluation of the chicken transcriptome by SAGE of B cells and the DT40 cell line. BMC Genomics 2004; 5(1):98.

Wang JH, Nichogiannopoulou A, Wu L, Sun L, Sharpe AH, Bigby M, Georgopoulos K. Selective defects in the development of the fetal and adult lymphoid system in mice with an Ikaros null mutation. Immunity 1996; 5(6):537-49.

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10(1):57-63.

Waters KM, Pounds JG, Thrall BD. Data merging for integrated microarray and proteomic analysis. Brief Funct Genomic Proteomic. 2006; 5(4):261-72.

Watson DJ, Crick FHC. A Structure for desoxyribose acid nucleic. Nature 1953; 171:737–738.

Webster G. Biologists Flirt with Models. Drug Discovery World Spring 2009; ddw.net-genie.co.uk.

Weill JC, Weller S, Reynaud CA. A bird's eye view on human B cells. Semin Immunol. 2004; 16:277–81.

Wellner RS, Pelayo R, Kincade PW. Evolving views on the geneology of B cells. Nat Rev Immunol. 2008; 8:95-106.

Wheeler DB, Carpenter AE, Sabatini DM. Cell microarrays and RNA interference chip away at gene function. Nat Genet. 2005; 37 Suppl:S25-30.

Wicker T, Robertson JS, Schulze SR, Feltus FA, Magrini V, Morrison JA, Mardis ER, Wilson RK, Peterson DG, Paterson AH, Ivarie R. The repetitive landscape of the chicken genome. Genome Res. 2005; 15:126-136.

Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HB, Saxild HH, Nielsen C, Brunak S, Knudsen S. A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol. 2002; 3(9):research0048.

Wu H, Kerr MK, Cui X, Churchill GA. MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments. In: Parmigiani G, Garrett ES, Irizarry RA, Zeger SL (eds).The Analysis of Gene Expression Data. London 2003: Springer.

Wu XL, Griffin KB, Garcia MD, et al. Census of orthologous genes and self-organizing maps of biologically relevant transcriptional patterns in chickens (Gallus gallus). Gene 2004; 340:213–25.

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 2005; 434(7031):338-45.

Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M, The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput Biol. 2007; 3(4):e59.

Yoshida T, Ng SY, Zuniga-Pflucker JC, Georgopoulos K. Early hematopoietic lineage restrictions directed by Ikaros. Nat Immunol. 2006; 7(4):382-91.

Yoshida T, Ng SY, Georgopoulos K. Awakening lineage potential by Ikaros-mediated transcriptional priming. Curr Opin Immunol. 2010; 22(2):154-60.

Zhao W, Serpedin E, Dougherty ER. Inferring gene regulatory networks from time series data using the minimum description length principle. Bioinformatics 2006; 22(17):2129-35.

Zhou XJ, Gibson G. Cross-species comparison of genome-wide expression patterns. Genome Biol. 2004; 5(7):232.