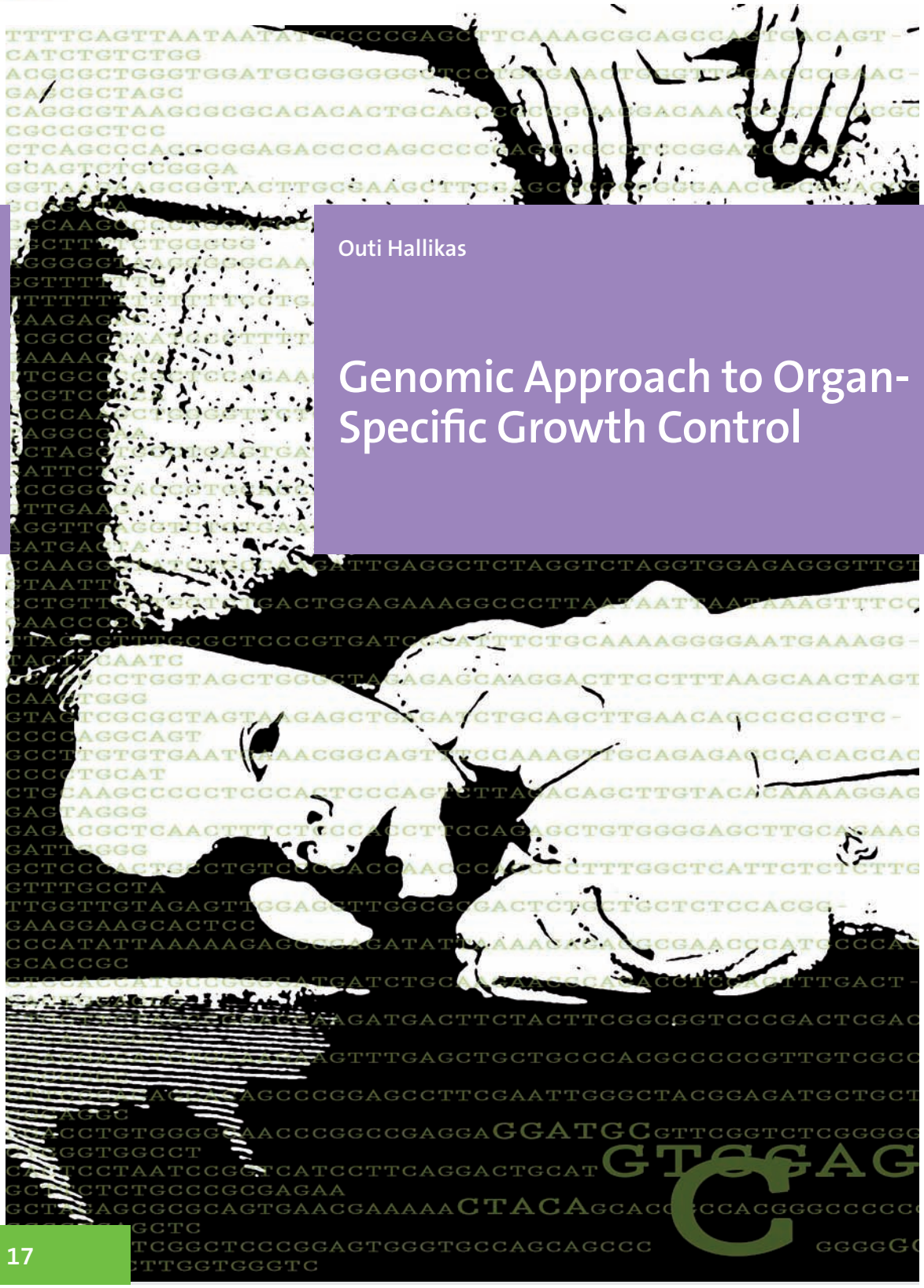
NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

RESEARCH

Outi Hallikas

Genomic Approach to Organ-Specific Growth Control



Outi Hallikas

GENOMIC APPROACH TO ORGAN-SPECIFIC GROWTH CONTROL

A C A D E M I C D I S S E R T A T I O N

To be presented with the permission of the Faculty of Biosciences, University of Helsinki, for public examination in Lecture Hall 2, Biomedicum Helsinki, on June 17th, 2009, at 12 o'clock noon.

National Institute for Health and Welfare, Helsinki, Finland

and

Division of Genetics, Department of Biological and Environmental Sciences, Faculty of Biosciences, University of Helsinki, Finland

RESEARCH 17

Helsinki 2009

ISBN 978-952-245-099-9
ISSN 1798-0054
ISBN 978-952-245-100-2 (pdf)
ISSN 1798-0062 (pdf)

Kannen kuva - cover graphic: photo by Elliot Yeung, graphic design by Merja Yeung

Helsinki University Print
Helsinki 2009

Supervised by

Academy Professor Jussi Taipale
Genome-Scale Biology Program
Institute of Biomedicine, University of Helsinki
Department of Molecular Medicine
National Institute for Health and Welfare
Helsinki, Finland
Karolinska Institutet, Department of Biosciences and Nutrition
Sweden

Reviewed by

Professor Jukka Jernvall
Developmental Biology Program
Institute of Biotechnology, University of Helsinki
Helsinki, Finland
Stony Brook University, NY, USA

Professor Jorma Palvimo
Institute of Biomedicine/Medical Biochemistry
University of Kuopio
Kuopio, Finland

Opponent

Dr. François Spitz
Developmental Biology Unit
European Molecular Biology Laboratory
Heidelberg, Germany

Custos

Professor Minna Nyström
Department of Biological and Environmental Sciences
Division of Genetics
University of Helsinki
Helsinki, Finland

Outi Hallikas, Genomic approach to organ-specific growth control
Publications of the National Institute for Health and Welfare, Research 17/2009, 122 Pages
ISBN 978-952-245-099-9; 978-952-245-100-2 (pdf)
ISSN 1798-0054; 1798-0062 (pdf)
http://www.thl.fi/

ABSTRACT

Growth is a fundamental aspect of life cycle of all organisms. Development of multicellular organisms requires generation of an individual with optimally sized, interdependent organs from a single cell. Body size varies highly in most animal groups, such as mammals. Moreover, growth of a multicellular organism is not uniform enlargement of size, but different body parts and organs grow to their characteristic sizes at different times. Currently very little is known about the molecular mechanisms governing this organ-specific growth.

The genome sequencing projects have provided complete genomic DNA sequences of several species over the past decade. The amount of genomic sequence information, including sequence variants within species, is constantly increasing. Based on the universal genetic code, we can make sense of this sequence information as far as it codes proteins. However, less is known about the molecular mechanisms that control expression of genes and about the variations in gene expression that underlie many pathological states in humans. This is caused in part by lack of information about the second genetic code that consists of the binding specificities of transcription factors and the combinatorial code by which transcription factor binding sites are assembled to form tissue-specific and/or ligand-regulated enhancer elements.

Our hypothesis is that growth *in vivo* is controlled by direct integration of ligand-regulated and tissue-specific transcription factor signals on enhancer elements of critical growth regulatory genes, such as the *Myc* genes.

This thesis presents a high-throughput assay for determining transcription factor binding specificities, which was then used to measure the DNA-binding profiles of transcription factors involved in growth control. We developed ‘enhancer element locator’, a computational tool, which can be used to predict functional enhancer elements. A genome-wide prediction of human and mouse enhancer elements generated a large database of enhancer elements. This database can be used to identify target genes of signaling pathways and to predict activated transcription factors based on changes in gene expression. Predictions validated in transgenic

mouse embryos revealed the presence of multiple tissue-specific enhancers in mouse *c-* and *N-Myc* genes, which has implications on organ specific growth control and tumor type specificity of oncogenes. Furthermore, we were able to locate a variation in a single nucleotide that carries a susceptibility to colorectal cancer to an enhancer element and propose a mechanism by which this single-nucleotide polymorphism might be involved in generation of colorectal cancer.

Keywords: organ-specific growth control, transcription factor, enhancer element, *c-Myc*, *N-Myc*, regulatory SNP

Outi Hallikas, Genomic approach to organ-specific growth control
Terveyden ja hyvinvoinnin laitoksen julkaisuja, Research x/2009, 122 sivua
ISBN 978-952-245-099-9; 978-952-245-100-2 (pdf)
ISSN 1798-0054; 1798-0062 (pdf)
http://www.thl.fi/

TIIVISTELMÄ

Kasvu on olennainen osa kaikkien eliöiden elinkaarta. Kaikki eläimet kasvavat, mutta niiden koko vaihtelee lajien välillä huomattavasti useimmissa eläinryhmissä, kuten esimerkiksi nisäkkäissä. Hedelmöittyneen munasolun kehittyessä monisoluiseksi yksilöksi kasvavat elimet oikeaan kokoonsa. On huomionarvoista, että monisoluisten eliöiden kasvaessa niiden koko ei suurene tasaisesti, vaan eri ruumiinosat ja elimet kasvavat lajityypillisen kokoisiksi kehityksen eri vaiheissa. Tällä hetkellä tiedetään hyvin vähän molekyylitason mekanismeista, jotka säätelevät elinkohtaista kasvua.

Viimeisen kymmenen vuoden aikana on sekvensoitu useiden lajien genomit. Genominlaajuisen sekvenssitiedon määrä kasvaa jatkuvasti, ja se sisältää tietoa myös lajien sisäisistä sekvenssivariaatioista. Universaalin geneettisen koodin perusteella pystymme lukemaan ja ymmärtämään saatavilla olevaa DNA-sekvenssitietoa niiltä osin kuin se koodaa proteiineja. Huomattavasti vähemmän tiedetään mekanismeista, jotka ohjaavat geenien luentaa DNA:sta RNA:ksi ja sairauksia aiheuttavista säätelyn variaatioista. Tämä johtuu osaltaan siitä, että tiedämme hyvin vähän niin kutsutusta ”toisesta geneettisestä koodista”. Toinen geneettinen koodi koostuu transkriptiotekijöiden tunnistesekvensseistä ja yhdistelmäkoodista, jonka mukaan transkriptiotekijöiden sitoutumispaikat muodostavat tehostajajaksoja, jotka säätelevät geeniluentaa.

Hypoteesimme on, että monisoluisen eliön kasvua säätelevät yhdessä kudoskohtaiset ja signaalireittien aktivoimat transkriptiotekijät. Näiden transkriptiotekijöiden välittämät aktivoivat ja inhiboivat viestit yhdistyvät transkriptiotekijöiden sitoutuessa tärkeiden kasvua säätelevien geenien, kuten *Myc*-geenien, tehostajajaksoihin.

Väitöskirjan tuloksissa kuvataan ensin tehoseulontamenetelmä transkriptiotekijöiden sitoutumisspesifisyyden mittaamiseen. Tätä menetelmää käyttäen määritettiin DNA-sitoutumisprofiilit useille kasvun säätelyssä keskeisille transkriptiotekijöille. Kehitimme myös tietokoneohjelman (enhancer element locator) tehostajajaksojen etsimiseen genomisesta DNA-sekvenssistä. Teimme genominlaajuisen ennusteen hiiren ja ihmisen tehostajajaksoista. Ennusteen tulokset koottiin tietokantaan, jota

voidaan käyttää signaalireittien kohdegeenien tunnistamiseen ja aktivoituneiden transkriptiotekijöiden ennustamiseen geenien ilmenemismuutosten perusteella. Kokeet muuntogeenisissä hiirissä osoittivat, että löysimme *c-Myc* ja *N-Myc* –geenien alueelta useita tehostajajaksoja, jotka ohjaavat geenien ilmenemisen tiettyyn kudokseen. Tulokset auttavat selittämään elinkohtaista kasvunsäätelyä ja onkogeenien kasvainspesifisyyttä. Lisäksi määritimme, että ihmisen paksunsuolen syöpään altistava yhden nukleotidin variaatio sijaitsee tehostajajaksossa ja esitimme mekanismin, jolla tämä yhden nukleotidin muutos voi vaikuttaa syövän syntyyn.

Avainsanat: kasvun elinkohtainen säätely, transkriptiotekijä, geenin säätelyelementti, *c-Myc*, *N-Myc*, säätely-SNP

CONTENTS

ABBREVIATIONS

Akt	v-akt murine thymoma viral oncogene homolog
Apc	ademtous polyposis coli
BAC	bacterial artificial chromosome
bp	base pair
3C	chromosome conformation capture
Cdk	cyclin-dependent kinase
CGNP	cerebellar granule neuron precursor
ChIP	chromatin immunoprecipitation
ChIP-chip	chromatin immunoprecipitation followed by DNA microarray
ChIP-seq	chromatin immunoprecipitation followed by parallel sequencing
CM	*cis*-module
c-myc, MYC	v-myc myelocytomatosis viral oncogene homolog (avian)
Ds	Dachsous
dPTEN	*Drosophila* phosphatase and tensin homolog
E	embryonic day
E box	enhancer box element
EEL	enhancer element locator
E2F	E2F transcription factor
Ets-1	erythroblast transformation specific or E twenty-six (avian erythroblastosis virus E26)

Ft	Fat
GFP	green fluorescent protein
GLI	glioma-associated oncogene
Hh	Hedgehog
HOX	homeobox
HUGO	The Human Genome Organization
IGF	insulin-like growth factor
kb	kilobase pair
lacZ	beta galactosidase reporter gene
L-myc, MYCL	v-myc myelocytomatosis viral oncogene homolog, lung derived (avian)
Max	MYC associated factor X
Mb	mega base pair
mRNA	messenger ribonucleic acid
N-myc, MYCN	v-myc myelocytomatosis viral related oncogene, neuroblastoma derived (avian)
$p27^{Kip1}$, CDKN1B	27 kDa cdk inhibitory protein, cyclin-dependant kinase inhibitor 1B
p190-B Rho-GAP, ArhGap5	Rho GTPase activating protein, Rho GTPase activating protein 5
PBM	protein binding microarray
PCR	polymerase chain reaction
PI3K	phosphaditylinositol 3-kinase
PN	postnatal day
pRB	retinoblastoma protein
SAGE	serial analysis of gene expression
SELEX	systematic evolution of ligands by exponential enrichment
Shh	sonic hedgehog

SNP	single-nucleotide polymorphism
S6k	ribosomal protein S6 kinase
Tcf4, TCF7L2	T-cell-specific transcription factor 4, transcription factor 7 -like 2
TFII	general transcription factor for RNA polymerase II
TGFβ	transforming growth factor beta
TOR	target of rapamycin
Wnt	wingless-type mouse mammary tumor integration site family
YAC	yeast artificial chromosome
YAP1	yes-associated protein 1

Where two abbreviations are given, the former is commonly used name and the latter is the name approved by the Human Genome Organization (HUGO).

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original articles referred to in the text by their Roman numerals:

I Hallikas O, and Taipale J. High-throughput assay for determining specificity and affinity of protein-DNA binding interactions. Nature Protocols 1, 215-222, 2006.

II Hallikas O*, Palin K*, Rautiainen R, Sinjushina N, Partanen J, Ukkonen E, and Taipale J. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell 124, 47-59, 2006.

III Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Björklund M, Wei G, Yan J, Niittymäki I, Mecklin J-P, Järvinen H, Ristimäki A, Di-Bernardo M, East P, Carvajal-Carmona L, Houlston RS, Tomlinson I, Palin K, Ukkonen E, Karhu A, Taipale J, and Aaltonen LA. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers ability to enhanced Wnt signalling. Nature Genetics, *In press*, 2009.

* These authors contributed equally.

III previously used in the Ph.D. thesis of Sari Tuupanen.

Author's contribution to the publications

I OH did all the experimental work. OH drafted the first version of the manuscript and wrote the manuscript with JT.

II OH performed the wet lab experiments either alone, or for the RNA in situ hybridizations in collaboration with NS and RR. OH also did some of the *in silico* analysis of genomic DNA. OH participated in writing of the manuscript with JT.

III OH performed transcription factor binding assay, mouse embryo work, and electrophoretic mobility shift assay. OH participated in the improving of the manuscript with other authors.

1 INTRODUCTION

Growth is one of the hallmarks of life. It is fundamental to all biological systems. Animals grow to distinctive sizes; their bodies and organs develop well-proportioned in a harmonious fashion. Giraffes have tall necks, elephants have long trunks, and never vice versa. Our arms grow to the same length and internal organs fit neatly into the body cavity. Abundant research has focused on defects and disarray of growth, namely cancer. But how is growth regulated when it is happening under normal circumstances, in an intact organism? How do we grow to be a certain size?

Animals reach characteristic sizes during fetal period and early life. Fish may continue to grow throughout life but most animals, like mammals and birds, cease to grow after reaching a certain size. Moreover, growth of a multicellular organism is not uniform enlargement of size, but various organs and body parts grow to their characteristic sizes at different times. Growth is a very basic aspect of development, yet physiological mechanisms that control growth remain poorly understood both at the level of animal and organ size (Conlon and Raff, 1999).

Based on the universal genetic code by which DNA encodes amino acids, we can make sense of the DNA sequence data as far as it encodes proteins (Crick, 1966; Nirenberg, 1963). Genomes of many species have been sequenced, including man (Lander et al., 2001; Venter et al., 2001), mouse (Waterston et al., 2002), and fruitfly (Adams et al., 2000), and the amount of genomic sequence data available is constantly increasing. The universal genetic code has allowed researchers to find new genes and estimate the total number of genes in the genomes. However, protein-coding sequence covers only a few percent of mammalian genomes. New codes and grammatical rules need to be resolved in order to understand the remaining genomic sequence. It is evident that genes are expressed in tightly controlled spatial and temporal patterns, but we do not know the code by which the expression is regulated. In this post-genomic era, the next big goal is to decipher the genetic code of regulation of gene expression.

Regulation of gene expression provides a potential, but currently unexplored, means of size control. This thesis will discuss size control and regulation of gene expression in mammals. The focus will be on molecular mechanisms of size regulation, with a special interest in organ size. However, as a lot of the research relating to size has come from animals

other than mammals, examples of those will also be mentioned where appropriate. Discussion on regulation of mammalian gene expression will bring enhancer elements into the spotlight.

2 REVIEW OF THE LITERATURE

2.1 Animal size

The largest mammal, and at the same time the largest animal ever lived, is believed to be the blue whale (*Balaenoptera musculus*). Its maximum-recorded weight is 190 tons, while the smallest bats and shrews weigh about 1.5 g. More than 100,000,000 Etruscan Shrews (*Suncus etruscus*), one of the smallest mammals today, are needed to balance the scale with one blue whale. The largest land-dwelling animal at present, the African bush elephant (*Loxodonta africana*), weighs as much as 3,000,000 shrews. Despite the difference in size, these organisms have the same overall body plan and the same organs, functioning in similar manner. Elephants and shrews had a common ancestor about 70 million years ago. Even though their genome projects are only on the way, they probably have highly similar coding sequences. In the mammalian genomes, there is extraordinary capacity to regulate size.

In addition to the observed size variation across species, animals within a single species hold a huge potential for size differences. Even though, in their natural environment, individuals of a single species seem relatively similar in size, application of artificial selection pressure for size reveals an ample genetic potential behind the uniform appearance. Various long-term breeding experiments and animal husbandry have produced size phenotypes in experimental and domestic animals. This has prepared the way for identification of the underlying genetic determinants. In an interesting study, chicken from a single founder population were bred and selected for size for 45 generations. This selection resulted in two lines of chicken showing 9-fold difference in size. Quantitative trait locus analysis revealed 13 loci affecting growth (Jacobsson et al., 2005). This result reflects the fact that body size is a classic example of a complex trait where multiple genes, as well as environmental factors, influence the final outcome. Growth rate and body size have also been studied in laboratory mice (Allan et al., 2005; Cheverud et al., 1996; Kenney-Hunt et al., 2006; Morris et al., 1999). These studies, similarly to the chicken experiments, show that it is very challenging to create a line of mammals or birds that is selected purely for size. Selection for size results in additional traits influencing level of energy intake, efficiency of energy conservation, fat deposition, appetite, reproductive capacity, physical activity, and so on. On one hand, feeding of

the high-weight chicken had to be restricted after selection age to avoid severe metabolic disorders. On the other hand, significant portion of the low weight chicken died of anorexia or never reached sexual maturity. These extreme cases highlight the importance of nutrition to growth and size (Jacobsson et al., 2005). Availability of food is a very significant factor influencing size of an animal. Even though the interest of this review is genetic control of growth, not environmental factors, the size-selected chicken show how interwoven the nutrition is also in genetic aspects of growth control.

The studies in mice and chicken have provided a large number of quantitative trait loci, which are linked to increase or decrease in body size and organ size (Kenney-Hunt et al., 2006; Park et al., 2006). However, the loci have been too numerous and the coverage of genetic markers has been too low to directly indicate candidate genes or other variants for further testing.

Well-documented regulators of overall size are growth hormone and its maybe most important downstream effector, the insulin-like growth factor (IGF) family of hormones. Deficiency or excess of these hormones results in perturbations of systemic growth, examples ranging from man to mouse and fly (e.g. Woods et al., 1996). The insulin-signal transduction pathway regulates several aspects of cellular physiology, including cellular growth, and uptake and utilization of glucose. There are good reasons to couple nutritional status with growth, as growth should occur in the presence of energy, and small size is more advantageous in the absence of resources.

The spectrum of dog breeds comprises a special case: one single species with enormous size variation. Dogs have been under man-made selection pressures for several thousands of years. The smallest and biggest dogs today may have 100 times difference in mass. Genetic information from individuals of different dog breeds was successfully used to locate a single *insulin-like growth factor 1* single nucleotide polymorphism (SNP) haplotype, which is common in small breeds and nearly absent in giant dog breeds (Sutter et al., 2007). Sutter and colleagues were not able to identify the causative variant. They found only one variation in the coding sequence, a synonymous SNP. However, they found numerous of small-breed-specific SNPs and other variations in introns and flanking genomic sequence (Sutter et al., 2007). Further studies will resolve the causative variant, good candidates being regulatory elements, which will be discussed in detail later.

2.2 Growth

Different sizes of animals are produced by differences in growth. Growth can be achieved by cell growth (increase in cell size), by increased rate of cell proliferation, by decreased rate of cell death, and by increase of extracellular space (accretionary growth) (Figure 1). In this study growth by increase of extracellular space will not be discussed.

In multicellular animals, the overall control of growth is crucial. The growth and proliferation within an organism, body parts, organs and individual cells has to happen in concert. Extracellular signals are important in this global control of growth. Growth factors stimulate cell growth. Mitogens stimulate cell division. Survival factors suppress cell death. Some factors may transmit combinations of these signals simultaneously. A single factor may function as a growth factor and a mitogen, for example.

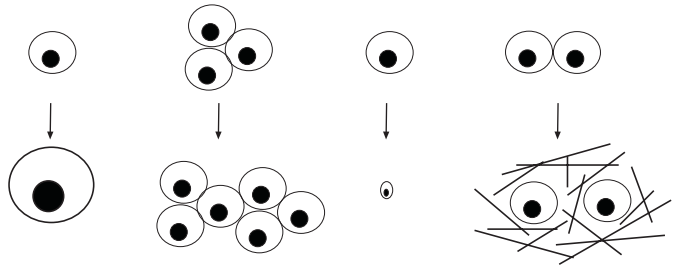


Figure 1. Growth can be regulated by change in cell size, in cell number, in rate of apoptosis, or in amount of extracellular matrix.

2.2.1 Cell size

Different aspects of growth are interdependent. However, cell growth, meaning enlargement of cells, is the first requirement for growth of an organism or organ. Proliferation, or division of cells, alone will never increase overall size, but just divide the existing mass into smaller and smaller units, the very phenomenon which takes place in cell divisions of mammalian morula.

Most mammalian cell types *in vivo* are of similar size in different species irrespective of the animal size; compared to a shrew, an elephant has more

of the same sized cells rather than bigger cells (Gregory, 2004; Stone et al., 1992). Yeast cells have a cell-size checkpoint, a mechanism that monitors cell size and allows cell division only after certain size is achieved (Fantes and Nurse, 1977). The size checkpoint is adaptable, however, as the size is dependent on specific culture conditions. There is an ongoing dispute over the existence of such size checkpoint in mammalian cells (Cooper, 2004; Echave et al., 2007). Definitely some specialized mammalian cells, such as neurons and muscle cells, can become very large and the size is species-specific (e.g. big animals have very long axons in their neurons). It is possible that different cell types in mammals have different control systems for size.

Increase in cell size may increase the size of an organism in flies. Mutations in the insulin/IGF pathway cause alters size of the fly mainly by altering growth (reviewed in Potter and Xu, 2001). *Drosophila melanogaster* fruit flies with a mutation in the *Drosophila phosphatase and tensin homolog* (*dPTEN*) gene are larger than normal due to bigger cell size. Conversely, inactivation of the *ribosomal protein S6 kinase* (*S6k*) gene in *Drosophila* results in smaller cells, and thus smaller fly overall. Members of the insulin/IGF family of ligands and receptors have been knocked out also in mice. As expected, deficiency of these proteins results in various degrees of size reduction, as well as other abnormalities (Rother and Accili, 2000). The insulin-like growth factor signaling is mediated downstream by phosphoinositide 3-kinase (PI3K), Akt and/or TOR pathway (Figure 2). Several members of this pathway give growth related phenotypes when knocked out in mouse. The most severe phenotypes are caused by PI3K catalytic subunit p110α or p110β deficiency, which lead to embryonic lethality at E10.5 and E3.5 respectively (Bi et al., 2002; Bi et al., 1999). Akt1 deficient mice are smaller than normal (Chen et al., 2001; Cho et al., 2001). Loss of S6k reduces mouse weigh by 10% (Dufresne et al., 2001). None of these mice was reported to have reduced cell size, even though deficiencies of the same pathway altered the cell size in the fly. Also, it was not addressed whether these effects could be caused by suboptimal function of placenta in the knockout mice. Genes affecting growth may also affect growth in the placenta, and a small/underdeveloped placenta may lead to smaller offspring.

Genetic manipulation of PI3K pathway in a specific organ of mouse, the heart, results in alteration of cell size as well as organ size (Shioi et al., 2000). Transgenic mice expressing constitutively active PI3K in the heart have larger heart size due to bigger cell size. Mice expressing dominant

negative mutants of PI3K have smaller hearts and smaller myocytes. Possibly this result relates to the plasticity of cell size of myocytes.

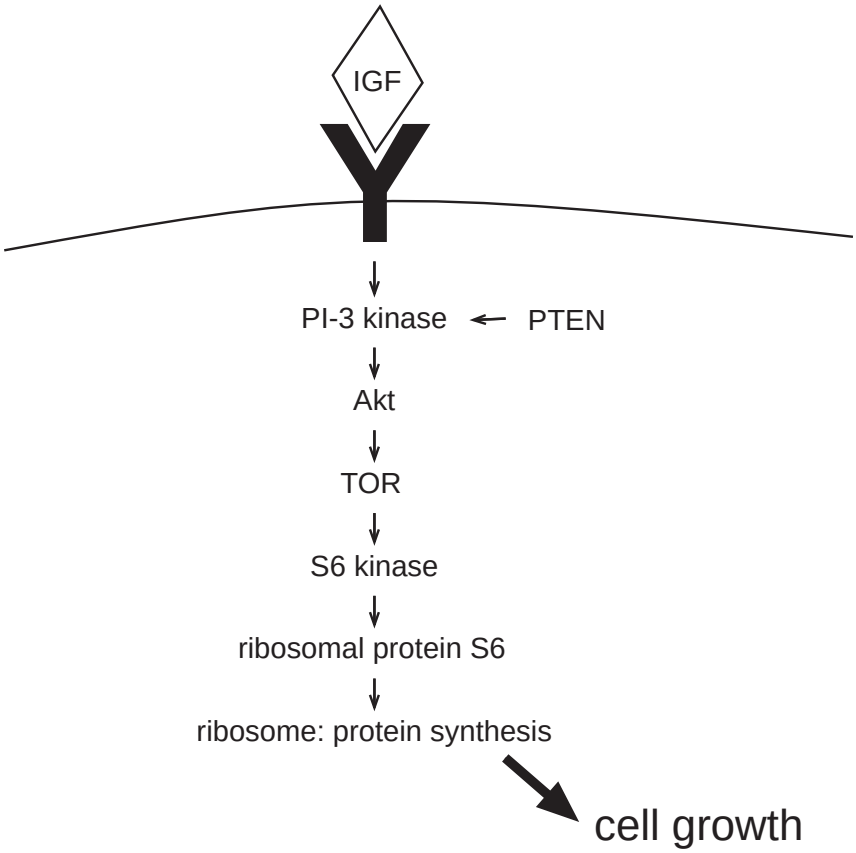


Figure 2. Schematic presentation of the IGF, PI3K, Akt and/or TOR pathway.

Mice lacking p190-B Rho-GAP (Rho GTPase activating protein) are reported to be about 30% smaller than normal mice and to die immediately after birth (Sordella et al., 2002). This is, to the best of my knowledge, the only reported case where the reduction in size of a knockout mouse is explained by smaller cell size than normal. p190-B Rho-GAP stimulates the GTP hydrolysis activity of Rho proteins. These mice also have smaller thymus than normal. It is proposed that p190-B Rho-GAP modulates

signaling from insulin/IGFs (Sordella et al., 2002). Why only this modulator of insulin/IGFs signaling would affect cell size and not the other components of the same pathway remains an open question.

As variation in cell size is fairly restricted in mammals, the cell number plays an important role in determining the final size. Cell number is principally affected by proliferation.

2.2.2 Proliferation

The cell division cycle, usually called the cell cycle, is the fundamental means by which all living cells propagate. It is of fundamental importance that the daughter cells receive identical, faithfully replicated chromosomes and all necessary cytoplasmic organelles. Therefore an elaborate control system is needed to coordinate the cycle as a whole. In multicellular organisms, in comparison to unicellular organisms, the control of cell cycle is even more vital as the cell proliferation has to be balanced within different cell types, tissues and organs of the organism. The cell cycle progression, as well as the cell size, of animal cells are controlled by signals from other cells (Conlon and Raff, 2003).

The progression of the cell cycle is thought to be dependent on the sequential activation of cyclin-dependent kinases (cdks), which depend on cyclin subunits for their activity. Oscillations in the synthesis and degradation of cyclins control the progressive steps of the cell cycle. Studies in yeast have identified Cdk4, Cdk6, pRB, the E2F family of transcription factors, Cdk2, cyclin D and cyclin E as critical regulators for cell cycle (Nurse, 1991).

Surprisingly, many of the cell cycle genes are largely dispensable for growth of a mouse. Mammalian cdks and cyclins, which were originally thought to be consequential for cell cycle, have one by one been shown replaceable in mouse development. Mice lacking Cdk4 (Rane et al., 2002; Tsutsui et al., 1999), Cdk6 (Malumbres et al., 2004), or Cdk2 (Berthet et al., 2003; Ortega et al., 2003) are viable. Moreover mouse lacking all interphase cdks, Cdk2, Cdk3, Cdk4 and Cdk6, develops till midgestation (Santamaria et al., 2007). Interestingly, genetic substitution of *Cdk1* by *Cdk2* causes embryonic lethality before E3.5 (Satyanarayana et al., 2008). Therefore, Cdk1 is necessary and sufficient for development of mouse embryos until E12.5. Only deficiency of the cyclins A2 and B1, which

activate Cdk1, cause very early embryonic lethality (Brandeis et al., 1998; Murphy et al., 1997).

Several other components of the cell cycle machinery have also been implicated to be critical in regulation of proliferation. Mice lacking cyclin dependent kinase inhibitor $p27^{Kip1}$ were reported to grow bigger than wild-type mice with an disproportionate enlargement of thymus, pituitary and adrenal glands, and gonads (Kiyokawa et al., 1996; Nakayama et al., 1996). Simultaneous targeted disruption of transcription factors E2F1, E2F2 and E2F3 in mice results in death before E9.5 (Wu et al., 2001). However, it is impossible to say to what extent these effects in cell cycle regulator deficient mice could result from placental effects. The lack of a protein may cause suboptimal growth in placenta and this may hinder the growth and development of the fetus. For example retinoblastoma knockout mice were originally found to be embryonic lethal at embryonic day 14.5 (E14.5), but with a wild-type placenta the mice develop to term (Wu et al., 2003).

2.2.3 Cell death

The net outcome in the cell number is a dynamic balance between cell proliferation and cell death. If apoptosis, or programmed cell death, is increased, the cell number is decreased, and if apoptosis is decreased the cell number rises, assuming that the rate of proliferation remains constant in a tissue.

Apoptosis is the suicide program by which cells can kill themselves when they are damaged or not needed (Raff, 1992). Activation of apoptosis occurs through specific signaling pathways and involves a caspase protease cascade. In apoptosis, the cell shrinks and condenses into fragments that are phagocytosed by macrophages and neighboring cells.

Apoptosis is a frequent phenomenon. It is estimated that in an adult human being 10 billion cells (0.01% of all cells in human body) die by apoptosis every day (Renehan et al., 2001). During development apoptosis is crucial. For example in the development of nervous system, the correct number of neurons is dependent on survival factors from the target tissue. In an adult organism, apoptosis works to maintain the correct size when cells are replaced due to damage or dysfunction. Mice lacking key apoptosis signal proteins exhibit, for example, brain overgrowth and interdigit webbing (Yoshida et al., 1998).

2.2.4 *Myc* genes

A classic growth regulator is *v-myc myelocytomatosis viral ongene homolog (avian)*, or by more familiar name known as *c-Myc*. *c-Myc* is a well-known proto-oncogene, and its expression is frequently altered in human cancers (Pelengaris et al., 2002). The c-Myc protein is a transcription factor that forms heterodimers with MYC associated factor X (Max) and regulates gene expression of a great number of genes via binding its consensus sequence, enhancer box element (E box). Generally c-Myc expression is required for proliferation of mammalian cells (Berns et al., 2000). c-Myc has been suggested to have a role in many growth-related aspects including proliferation, cell growth, differentiation and apoptosis. The Myc family of transcription factors has four additional members, N-Myc, L-Myc, S-Myc, and B-Myc. N-Myc and L-Myc are structurally and functionally similar to c-Myc, but their expression patterns are partially different (Hatton et al., 1996; Stanton et al., 1992). Significantly less is known about S-Myc and B-Myc.

Mice lacking c-Myc were originally reported lethal at E10.5 (Davis et al., 1993). In 2001, it was reported that, quoting the title of the paper, "c-Myc regulates mammalian body size controlling cell number but not cell size" (Trumpp et al., 2001). More recent studies have indicated, however, that regulation of size in mouse appears to be a placental effect, as c-Myc is required for branching morphogenesis of placenta (Dubois et al., 2008). Epiblast-restricted c-Myc null embryos with wild-type placenta die later than E10.5, but still before E12 (Dubois et al., 2008). The embryos also exhibit fetal liver hypoplasia, apoptosis of erythrocyte precursors, and functionally defective definitive hematopoietic stem/progenitor cells. Members of mammalian *Myc* gene family, *N-Myc* and *L-Myc* may take over *c-Myc* functions. Mice deficient in N-Myc die at E11.5 and have defects in multiple organs (Charron et al., 1992; Moens et al., 1992; Stanton et al., 1992). Placental rescue of *N-Myc* deficient mice is presently not available. Conditional knockout of *N-Myc* gene showed that it is essential for generation of nervous system (Knoepfler et al., 2002). Mice homozygous for a targeted null allele of *L-Myc* are viable, fertile and apparently healthy (Hatton et al., 1996). Conditional targeting of all the *Myc* homologs may be able to dissect the importance of *Myc* genes in more detail. Recent data indicates that *Myc* genes are involved in regulation of organ size, as N-Myc regulates proliferation in mouse retina and conditional deletion of *N-Myc* gene results in reduced retinal size (Martins et al., 2008).

2.3 Size of organs and body parts

Diversity of body parts in various shapes and sizes has created the multitude of mammalian species around us. The mechanisms that regulate size of mammalian organs and body parts are poorly understood.

2.3.1 Control of total cell mass rather than cell number

Several lines of evidence show that mammals, as well as many other animals, have mechanisms to control the total cell mass (the composite of cell size and cell number). Studies addressing total cell mass have exploited the fact that polyploidy increases cell size. For example polyploid salamanders, which have bigger cells, grow to the normal size (Conlon and Raff, 1999). The number of cells is reduced to compensate for increase in cell size, thus maintaining the normal size overall. There are numerous *Drosophila* mutants in which changes in proliferation lead to changes in cell size and maintenance of normal cell mass (Potter and Xu, 2001). Mammals are more sensitive to developmental perturbations than amphibians or insects, as tetraploid mice do not develop beyond E14.5. However, at that stage tetraploid mouse embryos contain about half of the normal cell number thus maintaining approximately the normal cell mass (Henery et al., 1992). Similarly, addition or reduction of cells into/from mouse morula does not change the final size of the mouse.

The same phenomenon, control of the total cell mass, has been observed at the organ level in several transplantation and partial organ excision experiments. If several fetal spleens are transplanted into a developing mouse, they will grow to reach the mass of one normal spleen (Metcalf, Transplantation, 1964). Following partial hepatectomy, liver cells proliferate to recover the original cell mass (Fausto et al., 2006). These observations show that size can be monitored and regulated in organ-specific manner, but presently many aspects of molecular mechanisms to assess and maintain total cell mass are unknown.

Processes to maintain the normal size include ‘competition’ and ‘compensation’. These have been studied particularly by in *Drosophila* imaginal discs, where the faster-growing cells outcompete the slower-growing or missing cells (Nijhout and Emlen, 1998). Fly cells of an organ also compensate their size to correct an excessive of deficient number of cells. Competition has been proposed to function also in mammals (Oliver et al., 2004).

2.3.2 Tissue-specific growth factors

A conceptually simple mechanism to facilitate organ-specific growth is the use of tissue-specific growth factors, which would stimulate or inhibit growth in specific target tissue. Myostatin, a transforming growth factor β (TGFβ) family member, inhibits the proliferation of myoblasts that fuse to form skeletal muscle cells. Deletion of myostatin gene in mice makes the muscles grow larger than normal (McPherron et al., 1997). Two breeds of cattle, bred for beef production, also have mutations in myostatin gene (McPherron and Lee, 1997).

2.3.3 Regulative versus autonomous growth

Even though our understanding of organ size determination is limited, some themes have arisen from the present knowledge. The control of cell growth and proliferation within an organ may be autonomous or regulative(Stanger, 2008). An organ under autonomous growth control maintains the size information and will not change size due to external signals or perturbations. In contrast, an organ that grows in regulative fashion, obtains signals from outside the organ and the growth is controlled at the level of the whole organism. An example of autonomous growth regulation is the pancreas: experimental reduction of fetal pancreatic progenitor cells results in a smaller pancreas (Stanger et al., 2007). Pancreatic progenitor cells, as early as E9.5, are confined to form a limited part of the future pancreas. No compensation of pancreatic size occurs after ablation of progenitor cells. The information about the organ size is contained locally within the cells of the fetal organ, rather than in the system of the organ or organism. Conversely, two-thirds of fetal liver could be removed in early gestation, and after 4 days its mass was approximately back to normal (Stanger et al., 2007). Liver maintains this regenerative capacity in adult organism as well. Based on this, there seems to be two types of organs. The first type depends on intrinsic signals for growth and does not regenerate in fetal life nor adulthood. The second type depends on extrinsic signals for growth, does regenerate in adulthood and exhibits regulated growth in development. Organs that have the capacity to replenish themselves to varying extent throughout adult life are the intestine, blood, skin and liver. The kidney, spleen and pancreas are examples of organs that do not readily regenerate in full-grown organisms.

The above study suggests that the pancreatic size is determined by number of progenitor cells. However, it seems that different mechanisms function in determining the size of the very same organ. Conditional removal of *adenomatous polyposis coli* (*Apc*) gene in pancreatic epithelial cells from E10.5 onwards causes pancreas to enlarge several fold due to hyperplacia of pancreatic exocrine cells (Strom et al., 2007). Conditional inactivation of *c-myc* gene in pancreata lacking *Apc* completely reversed the hyperplacia (Strom et al., 2007).

2.3.4 Patterning

Size of body parts is coupled to pattern formation. Patterning is an embryologic process that specifies different cell fates spatially and temporally in a structure that is initially largely homogenous. Defects in patterning result in altered shape, size, or complete omission of the effected organs or body parts. On the one hand, if patterning is perturbed, part of a structure might be missing and result in reduced size. On the other hand, ectopic expression of a signaling molecule may result in pattern duplications and excessive growth. A drastic example of perturbed patterning was caused by thalidomide, a sedative, which was used in early pregnancy and resulted in congenital defects of limbs in thousands of children (Knobloch et al., 2007).

Hox genes, which determine body regions along the anterior-posterior axis, are likely to affect size of homologous structures, e.g. vertebrate ribs, within an individual. Experimental evidence for this has been found in *Drosophila*, where Hox gene *Ultrabithorax* limits size of the haltere (small appendage used to balance during flight) by restricting transcription and mobility of the morphogen Decapentaplegic (Crickmore and Mann, 2006).

Patterning has been studied particularly in limb development. Recent evidence indicates that proliferation and pattern formation may be regulated by the same signaling pathways. Alterations in the duration and range of signaling may underlie morphological differences in the evolution of vertebrate limbs (Yang, 2009).

2.3.5 The steepness hypothesis of morphogen gradients

The Hippo signaling pathway has been identified in *Drosophila melanogaster* as an intrinsic mechanism that restricts organ size in

development (Edgar, 2006). The Hippo pathway was also found to regulate liver size in mice: Overexpression of a constitutively activated Hippo pathway target, transcriptional coactivator yes-associated protein 1 (YAP1), in mouse liver causes reversible 4-fold increase in liver size (Camargo et al., 2007; Dong et al., 2007). The Hippo pathway may serve as the signal to stop tissue growth once the correct size has been achieved in liver. A mechanistic hypothesis, the steepness hypothesis, has been proposed to link morphogen gradients to cell polarity and growth in *Drosophila* (Lawrence et al., 2008). Lawrence and colleagues propose that morphogens, which set up the pattern of the organ, determine a linear gradient of protocadherins, Fat and Dachsous (Ft and Ds). According to the steepness hypothesis, the direction of the Ds/Ft gradient determines the cell polarity within an organ and the steepness of the same gradient determines the die-or-divide decisions through the Hippo pathway (Lawrence et al., 2008). Even though the hypothesis is still incomplete and there are many open questions remaining, it is an inspiring framework, because it could explain how the information about growth decisions and dimensions of an organ is combined to create the correct size.

Vertebrates have four Ft homologs and two Ds homologs (Rock et al., 2005). Fat4 has recently been knocked out in mouse (Saburi et al., 2008). The $Fat4^{-/-}$ mice die at birth; they have curved body axis and curly tails. They have defects in the inner ear, similarly to other vertebrate mutants of core planar cell polarity genes. Interestingly, loss of Fat4 disorganizes the oriented cell divisions and tubule elongation in kidney development.

2.4 Overview of regulation of mammalian gene expression

As coding regions are similar between different metazoan species, it is likely that organ and animal size are encoded in the more divergent regulatory regions that control gene expression. Even though the comprehensive data is not available today, it seems that metazoans, ranging from sea squirt to squirrel, have similar numbers of genes coding for proteins and RNAs (Ensembl database, www.ensembl.org). There is multiple evidence for mutations causing changes in amino acid sequence that result in altered physiology or morphology (for example fur color Majerus and Mundy, 2003). However, reported changes in protein coding sequences resulting in phenotypic alteration are rare relative to the extensive diversity of body forms (Carroll, 2005; King and Wilson, 1975).

To the best of my knowledge, presently there is no report of change in amino acid sequence causing alteration of size between species. Moreover, many of the developmental regulators, which affect for example growth, are highly conserved across species. A considerable part of the phenotypic diversity between metazoan species seems to rise from somewhere other than difference in the number of genes or variation in their coding sequences. A very good candidate for additional origin of morphological diversity is regulation of gene expression.

Mammalian gene expression is controlled at many levels. The first and most general level of control is the chromatin state. The second level of control is initiation of transcription. Further levels of controlling gene expression include processing of the transcript, transport of messenger RNA (mRNA) into cytoplasm, control of translation, degradation of mRNA, and ultimately control of protein activity and degradation. It is not possible to exhaustively evaluate the relative importance of these different levels of control. The most important regulation in eukaryotes has been considered to take place at initiation of transcription. Even though evidence is increasing for important examples of control at other levels, initiation of transcription is still the starting point, without which the regulation at further levels could not happen.

In recent years, the non-coding RNAs have gained much attention. MicroRNAs also perform functions regulating transcription following many common (and some different) principles as transcription factor-mediated gene regulation (Hobert, 2008). They modify existing transcriptional programs by channeling mRNAs into degradation. Their importance lies in the fine-tuning and pruning the expression patterns rather than generating new patterns. In this discussion, I will focus on regulation of gene expression at the level of transcriptional initiation by distal enhancer elements.

2.4.1 Chromatin state

The nucleus is a complex functional unit. Even though it is not divided into membrane-bound compartments like cytoplasm, it contains many functional sections involved in different aspects of transcription. Genomic DNA is distributed into spatially and functionally distinct entities within the nucleus. Chromosomes occupy specific nuclear spaces or territories, and genetic loci are located in different types of chromatin (Carmo-Fonseca, 2002). Moreover, the position of a given locus within the nucleus correlates

with its expression. Transcriptionally active genes tend to localize to the periphery of the nucleus in the ”transcription factories” (Kosak and Groudine, 2004). Co-expressed genes co-localize in the same ”factory” (Osborne et al., 2004).

When DNA is highly condensed, it is thought that transcription factors cannot access their binding sites, which are packaged around histones and in larger chromatin structures. This feature contrasts the two basic types of chromatin: euchromatin is transcriptionally active and less condensed, while heterochromatin is highly packaged, transcriptionally silent, and inaccessible to DNA-binding proteins and chromatin-modifying enzymes. Various mechanisms are exploited to control the access of eukaryotic transcription factors to the binding sites buried in chromatin. Histone modifications, such as methylation, acetylation, ubiquitination, phosphorylation, and sumolation of specific recidues, change DNA-histone interactions. Modifications in histones are recognized by other proteins and may directly influence higher–order chromatin structure. Histone variants, which differ from canonical histones mainly in the histone tails, affect different stages of transcription. However, it appears that DNA packaged around nucleosomes is still to some extent accessible to DNA-binding proteins (Li et al., 2007). A prediction model suggests that nucleosome density at promoters is also lower than elsewhere in the genome (Richmond, 2006).

Nuclear compartments and chromatin states are often dynamic; chromatin can change form and genetic loci may shift between active and silent chromatin. However, some chromatin signatures are inherited, resulting in epigenetic inheritance. Recent work gives genome-wide evidence that histone modifications at promoters are largely invariant in different cell types, while enhancers are marked with highly cell-type specific histone modification patterns (Heintzman et al., 2009).

2.4.2 Eukaryotic transcriptional machinery

The critical player of the eukaryotic transcriptional machinery is RNA polymerase II, which transcribes all protein-coding and small nucleolar RNA genes, as well as most microRNA and small nuclear RNA genes. In distinction from its prokaryotic counterpart, eukaryotic RNA polymerase II requires a repertoire of additional protein factors to function. General transcription factors for RNA polymerase II (TFIIs), such as TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, are involved in the initiation of

transcription. RNA polymerase II and the general transcription factors form the preinitiation complex and can direct transcription at basal level.

In multicellular organisms, there is further complexity of transcription apparatus compared to unicellular eukaryotes (Levine and Tjian, 2003). Tissue-specific TATA binding protein-associated factors (TAFs) of the core promoter complex permit selective activation of genes in a particular tissue (Freiman et al., 2001). The mediating factors and different compositions of core transcription apparatus provide variability in the possible assemblies of transcriptional apparatus. Multisubunit cofactor complexes (equivalent to the yeast mediator) connect distal activators to the core complex and may be induced to undergo conformational changes to activate transcription. Cofactor complexes have diversified extensively in metazoans (Glass and Rosenfeld, 2000; Malik and Roeder, 2000). Diversification of cofactor complexes may reflect the increase in complexity of distal transcriptional regulation. Chromatin remodeling and modifying complexes assist transcription apparatus to travel through chromatin. There is also evidence that these complexes have diversified considerably in metazoan and mammalian evolution (Wang et al., 1996). More factors of the transcription apparatus may surface with further research.

2.4.3 Regulatory elements of genomic DNA

Various transcriptional regulatory elements in genomic DNA can be subdivided into categories: core promoters, proximal promoter elements, enhancer elements, silencers, insulators and global control regions/locus control regions (Maston et al., 2006) (Figure 3).

Core promoter

The core promoter is approximately 30-75 base pairs long docking site for the preinitiation complex. It defines the transcription start site and direction of transcription. Mammalian promoters may contain various conserved sequence elements, the most famous being the TATA box. Mammalian promoters can be divided into two classes: conserved TATA box-enriched promoters, which initiate transcription at a well-defined site, and more plastic, broad and evolvable CpG-rich promoters (Carninci et al., 2006). Variability of core promoters has functional significance as composition of

the promoter determines its responsiveness to specific activating and inactivating signals (Morris et al., 2004; Smale and Kadonaga, 2003).

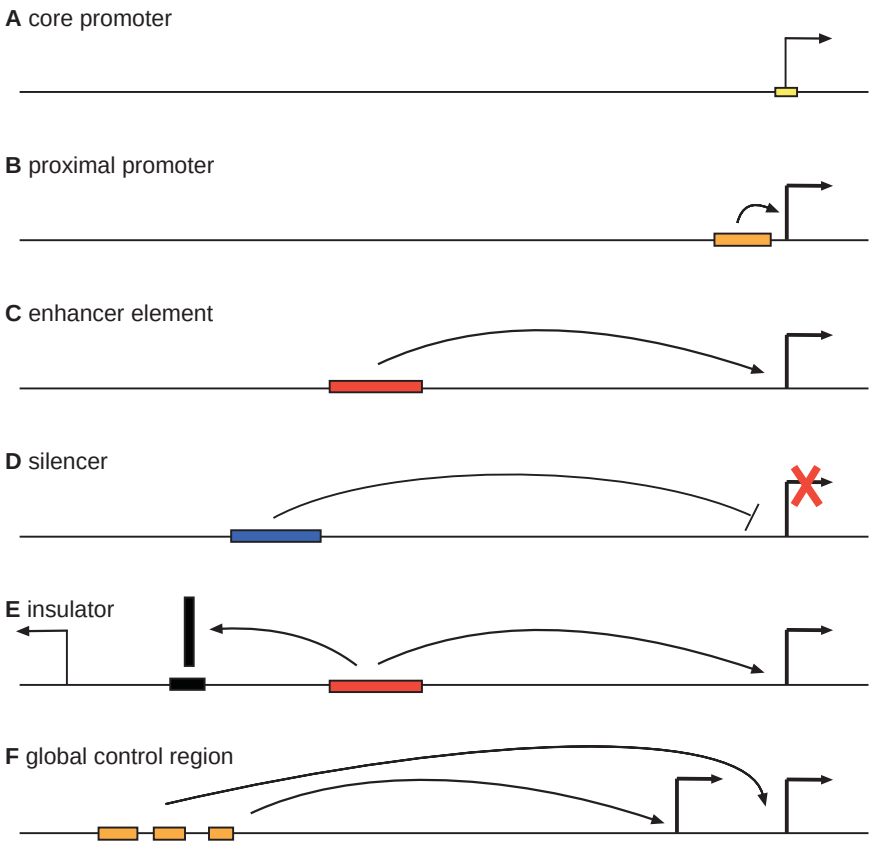


Figure 3. Transcriptional regulatory elements.

Proximal promoter elements

The region immediately upstream of the core promoter typically contains a lot of transcription factor binding sites. Historically, most regulatory sequences have been found within a few kilobase (kb) area upstream from the transcription start site. This area, called proximal promoter, can often recapitulate some aspects of gene's expression. The proximal promoter contains clustering of transcription factor binding sites. Transcription

factors, which bind to the proximal promoter, may provide binding sites for activating or suppressing transcription factors of distal enhancer elements connecting enhancer to the core promoter (for example Perkins et al., 1996).

Enhancer elements

Enhancer elements are genomic DNA sequence elements that have ability to regulate transcription irrespective of their orientation in relation to the target promoter. Typically enhancer elements specify gene expression quantitatively, spatially and temporally to a particular tissue or cell type at specific developmental stage. They range in size from hundred base pairs to a few kilobases. Enhancer elements may be upstream or downstream hundreds of kilobases, or even a megabase, away from the target promoter. An enhancer element enforces the regulation from across genes, or an enhancer may be located in another gene's intron (Lettice et al., 2003). Also, interchromosomal activation has been suggested in the regulation of interferon-γ expression in mouse (Spilianakis et al., 2005).

Silencers

Silencers are composed of binding sites for transcription factors, all or some of which have repressor activity. Silencers function in similar way as the enhancers; they are "negative enhancer elements". Repressor function of silencers requires recruitment of repressing transcription factors or cofactors. Repressor function may take place via blocking a binding site from an activating transcription factor, inhibiting the transcription apparatus, masking the activation surface of an activator, recruitment of chromatin remodeling complexes, or recruitment of histone modifying enzymes.

It was previously thought that eukaryotic transcription is silenced at default state, which would make silencers largely gratuitous in the regulation of transcription. This view has been challenged by reports of human genome being pervasively transcribed (Birney et al., 2007). Silencers do not seem to be completely unnecessary in control of eukaryotic transcription, as some silencers have indeed been found in eukaryotic genomes. An example of disruption of silencer function in humans is fascioscapulohumeral muscular dystrophy. The patients have a deletion of a chromosomal repeat, and that abolishes a binding site of a repressor complex (Gabellini et al., 2002).

Insulators

Insulators are 0.5-3 kb long sequence elements that limit action of transcriptional regulatory elements. Insulators come in two varieties: those that block communication between enhancer and promoter when located between them, and those that restrict the spread of heterochromatin. In vertebrates, there is only one known insulator protein, CTCF (CCCTC-binding factor) that binds specific sequence element, blocks enhancer activity and also restricts heterochromatin. CTCF was originally discovered as a repressor of the chicken *c-Myc* locus (Lobanenkov et al., 1990). Since then CTCF has been found to function in the regulation of several genes (Wallace and Felsenfeld, 2007).

Insulator elements are one of the mechanisms ensuring that the right enhancer will interact with the right promoter. The exact mechanism of insulator function is still an open question (Bushey et al., 2008). According to a prominent model, the loop domain model, insulators form chromatin loops and are involved in the nuclear organization of chromatin (Bushey et al., 2008).

Global control regions/Locus control regions

In mammalian genomes a few cases have been found where gene regulatory elements act on a number of genes, for example *β-globin* and *Hox* gene clusters. The regulatory interactions involving many partners may be quite complex. There are 'global enhancers' that modulate expression of several genes (Spitz and Duboule, 2008). 'Global control regions', or 'locus control regions', are composed of groups of regulatory elements that act on multiple genes (Spitz et al., 2003). The regulatory regions and their target genes form regulatory landscapes on the genomic DNA, which may vary in different tissues.

All the regulatory elements mentioned above are regions of genomic DNA sequence that in cellular context can have important effects on transcription. These DNA sequences cannot deliver such an effect alone; a functional party is needed. Transcription factors bound to the enhancer elements exert the effects of promoting, enhancing or silencing of transcription.

2.5 Transcription factors

Transcription factors are DNA-binding proteins that regulate transcription. Transcription factors are composed of two domains: the DNA-binding domain and the activating (or inactivating) domain. The DNA-binding domain determines the binding specificity. Transcription factors have affinity for short, usually 5-20 base-pair DNA sequences. The binding specificity is for degenerate sequences, i.e. only some positions are critical and some allow flexibility in base composition. The activating domain functions independently of the DNA-binding domain. The fundamental activating function of transcription factors is to attract, position and modify the general transcription factors, cofactor complex and RNA polymerase II to boost initiation of transcription. This can happen either directly via protein-protein interactions to the components of the transcription machinery, or indirectly via altering the chromatin structure. Alterations of chromatin structure involve local modifications by various chromatin remodeling complexes, histone chaperones and histone removing enzymes (Perillo et al., 2008).

Transcription factors are divided into families based on their DNA-binding domains. There are dozens of families of DNA-binding proteins, and only few families are mentioned here. GLI/Ci transcription factors that are activated by the Hedgehog (Hh) signaling pathway belong to the zinc finger family of transcription factors (Pavletich and Pabo, 1993). Tcf4 is activated by the Wnt (wingless-type mouse mammary tumor integration site family) pathway and contains the high mobility group box as defining domain (Poy et al., 2001). c-Ets1 and other Ets transcription factors belong to a family of their own, Ets family (Oikawa and Yamada, 2003). Myc transcription factors bind DNA as heterodimers with Max protein. They belong to a large family and have basic-helix-loop-helix-leucine zipper domain that combines two functions: dimerization and DNA binding (Nair and Burley, 2003).

2.5.1 Determining the binding specificity of transcription factors

A crucial aspect of studying transcriptional regulation via enhancers is the binding specificity of transcription factors. Different methods exist for determining the DNA-binding affinity of proteins. DNA-binding specificity is typically presented as positional weight matrix, which is amenable to computerized sequence analysis (Table 1). The quality of positional weight

matrices is critical to success in enhancer prediction by large-scale genomic sequence analysis. Even small distortions in the weight matrices lead to erroneous result due to the inherently degenerate nature of transcription factor binding specificities and the megabase scale of DNA sequence analyzed.

Table 1. Example of positional weight matrix (transcription factor GLI2).

	Binding site nucleotide position										
	1	2	3	4	5	6	7	8	9	10	14
A	0.01	0.76	0.03	0.00	.081	0.00	0.17	0.03	0.96	0.37	0.04
C	0.00	0.14	0.97	1.00	0.07	1.00	0.82	0.94	0.01	0.48	0.06
G	0.95	0.11	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.11	0.84
T	0.04	0.00	0.00	0.00	0.09	0.00	0.01	0.03	0.02	0.04	0.07

There are two basic lines of approach that have been used to generate positional weight matrices: 1) analysis of occupied sites in given DNA, and 2) analysis of affinity of a transcription factor to DNA. Examples of the first include DNAse footprinting, coimmunoprecipitation of genomic DNA with a transcription factor, and modern techniques of chromatin immunoprecipitation followed by sequencing of cloned fragments (Wei et al., 2006), genomic microarray (ChIP-chip) (Carroll et al., 2005), or parallel sequencing (ChIP-seq) (Robertson et al., 2007). These methods determine occupied sites in a specific DNA sample and the sites are aligned to deduce binding specificity. Chromatin immunoprecipitation based methods generally give about 100-500 bp sequence reads around the binding site and require use of motif discovery algorithm to find the binding motif. As these methods do not directly measure affinity, they are prone to statistical error, when small numbers of binding sites are aligned, as is the case with older methods. ChIP-chip and ChIP-seq methods identify such large numbers of sequences that the methods circumvent most of the pitfalls of traditional alignment-based methods. These experiments still may fail to identify significant binding sites, because the studied transcription factor may not bind all sites under the specific conditions. Also, ChIP-chip and ChIP-seq

techniques may identify binding sequences for associated factors rather than the intended transcription factor. The special value of ChIP-chip and ChIP-seq methods lies in their ability to provide important data on occupied sites in a particular cell line or tissue sample.

The DNA affinity of a transcription factor can be obtained by binding the protein of interest to every permutation of n positions of duplex DNA. The output can be read by determining the intensity of binding or sequencing sufficient number of bound DNAs. Earlier applications of this idea are SELEX (Systemic Evolution of Ligands by Exponential enrichment) and SELEX-SAGE (SELEX-Serial Analysis of Gene Expression) -type methods, but these techniques are labour-intensive in large scale as they include multiple rounds of PCR, cloning and sequencing (Roulet et al., 2002). High throughput has been achieved using bacterial one-hybrid system (Meng et al., 2005; Noyes et al., 2008). However, possible incorrect folding of mammalian DNA-binding domains when expressed in *E. coli* will be a limitation of this method. Protein binding microarrays (PBMs) utilize binding of labeled protein on microarray containing every permutation of up to ten positions of duplex DNA (Badis et al., 2008; Berger et al., 2006; Warren et al., 2006; Zhu et al., 2009). These studies provide high quality data for *in silico* identification of binding sites. Despite the several methods available, the knowledge of binding specificities is still presently lacking for hundreds of mammalian DNA-binding proteins.

It would be of great value to be able to predict transcription factor binding specificity from the amino acid sequence of the protein. Limited success in binding specificity prediction is probably caused by large influence of secondary and tertiary structure of DNA-binding domain on binding specificity and interdependence between positions of the binding site further complicating the situation (Benos et al., 2002; Miller and Pabo, 2001; Wolfe et al., 2001). The best predictions so far are based on existence of knowledge on binding specificity of a sufficiently similar homolog or ortholog (Berger et al., 2008).

Besides DNA affinity, there are several variables affecting the functionality of transcription factor binding sites *in vivo*: concentrations of transcription factors in the specific cell, secondary interactions with other proteins (complex formation) and chromatin environment of the specific cell. It is likely that transcription factors, or at least some of them, are in limiting concentration in the cell.

2.6 Enhancer elements

What is known about enhancer elements currently, and how have they been studied?

2.6.1 Physical contact between enhancer and promoter

Regulatory elements in mammalian genomes are widely spread out. Distal enhancer elements may be located hundreds of kilobases away from the transcription start site. How is the signal from the enhancer element passed on to the initiation of transcription? The prevailing model is that DNA looping allows the enhancer element and the promoter to interact forming a protein complex. This complex consists of transcription factors binding to the enhancer, the transcription machinery and possibly some connecting factors (West and Fraser, 2005) (Figure 4).

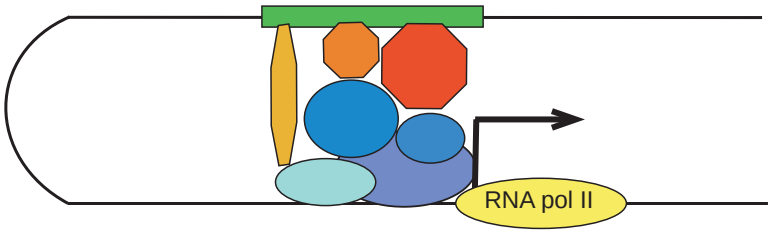


Figure 4. Looping model. Protein-protein interactions between the enhancer element-bound transcription factors (brown) and general transcription factors (blue) stabilize or activate the basal transcription machinery to initiate transcription.

It has been challenging to study how the activating signal from an enhancer is passed on to the promoter. Microscopy and chromosome conformation capture (3C) have provided data to show that DNA looping brings enhancer element and its target promoter in close proximity (Dekker et al., 2002). Such interactions are, however, transient and overall chromatin state in the nucleus is always dynamic and variable. This complicates the interpretation of results. Various models have been proposed on how the enhancer and promoter come to contact. Passive diffusion and different active processes, such as enhancer tracking along the chromatin, movement in actin-dependent fashion across nucleus and active bending of chromatin, have been proposed as possible mechanisms (Dekker, 2008; Perillo et al., 2008).

2.6.2 Combinatorial code of enhancer elements

An enhancer element is composed of a cluster of transcription factor binding sites. The precise order and organization of the sites is probably important for enhancer function, but currently limited information exists about the rules of enhancer composition. The only principles governing the orientation and spacing of transcription factors known today concern a few transcription factors which bind DNA as homodimers and heterodimers (Remenyi et al., 2004).

Biological evidence suggests a combinatorial code of transcriptional regulation to allow the vast variety of expression patterns observed within a single individual and the variety seen across species. Combinations of transcription factors could produce a large enough coding potential to explain the observed variety in phenotypes. It is estimated that the human genome contains about 1400 genetic loci encoding proteins that bind DNA in a sequence specific manner (Vaquerizas et al., 2009). As there are about 30,000 genes in the human genome, and most of them are expressed at multiple developmental time points and places, the number of transcription factors acting individually is not sufficient to create the coding potential for the variety of expression patterns required across the human lifespan. Moreover the nearly same proteins as in human being are expressed in chimpanzee resulting in considerably different kind of animal. This is indirect evidence for existence of combinatorial code by which transcription factor binding sites are assembled to form tissue-specific enhancers. However, our knowledge of the "grammar" of the code is presently very limited.

2.6.3 Conservation of enhancer elements in different mammalian species

Even though enhancer elements are mainly located in non-coding genomic sequence, their sequences are thought to be evolutionary conserved in order to maintain their function.

Mutations are a source of variation. Mutations with greater pleiotrophic effects are expected to have more deleterious effects on organism's fitness than mutations with less widespread effects. Enhancer elements can be thought to be a more fertile ground for evolution than amino acid coding regions, because activity of an enhancer is often restricted to certain tissue and time point, therefore it is not pleiotropic. Furthermore, redundancy

(which can be achieved by multiple enhancers overlapping in function) and compartmentation (enhancers driving expression in specific tissues) increase "evolvability", the capacity to generate tolerable, inheritable variation (Gerhart and Kirschner, 2007).

2.6.4 Genome-wide prediction of enhancer elements

A plethora of computer programs has been designed for prediction of enhancer elements during this millennium alone. They often follow some similar, but also varying, principles. Commonly used principles include the following:

1) Conservation of, particularly non-coding, sequence in different species is generally considered to imply functional importance. Many programs use local or global DNA sequence alignment of two or several species of varying evolutionary distance to detect the evolutionary conservation (Woolfe et al., 2005).

2) Regulatory elements are expected to have clustering of binding sites of certain affinity. A variety of pattern recognition programs has been used for locating binding sites. Binding motifs from different sources, such as the Transcfac database (www.biobase.de) or the Jaspar database (www.jaspar.cgb.ki.se), or researchers' own results, are utilized in the pattern searching (Knuppel et al., 1994; Sandelin et al., 2004). Some programs aim to find several binding sites for a single or few transcription factors within one enhancer element (Rajewsky et al., 2002).

3) Expression profiling experiments potentially identify groups of coexpressed genes. Some approaches use such sets of genes to find similar motifs, which potentially could be regulatory elements (Kloster et al., 2005).

4) Some programs use the assumption that regulatory elements have some characteristic properties, which software can be trained to recognize. This has given rise to machine learning programs (Kolbe et al., 2004).

All the above search approaches appear individually and in varying combinations in literature. The amount of genomic sequence analyzed in various approaches is also highly variable. Many studies have been limited to the promoter proximal region (Suzuki et al., 2004). It is typical to search for binding sites in the sequences that were found to be homologous by

alignment. Another typical variation is to use motif discovery algorithms for the aligned sequences.

An extreme approach for searching regulatory elements has been identification of ultraconserved sequences (Woolfe et al., 2005). Four elements (ranging in length between 222-731 bp,) which were 100% conserved in man, mouse and rat were deleted in mice without observable consequences to the mice (Ahituv et al., 2007). Even though conservation is considered to reflect functional importance, ultraconservation alone gives no hint of the kind of function that the sequence might have. These sequences might function as for example replication origins or modulators of chromatin structure. Also, we do not know the prevalence of redundancy in such elements. The ENCODE project has attempted to locate regulatory elements in subset of human genome. They conclude that sequence conservation alone is not sufficient to locate regulatory elements (Birney et al., 2007; McGaughey et al., 2008).

2.6.5 Experimental approaches for identifying tissue-specific enhancer elements

Whereas computer software can be used successfully to predict large numbers of potential regulatory elements, the experimental validation of predictions in large scale is presently challenging.

Transient transgenics in mouse embryos

Producing transient transgenic mouse embryos carrying the enhancer element and a reporter gene (*lacZ* or *GFP*) can be used to biologically detect the activity of the predicted mammalian enhancer elements. The limitations of this procedure include relatively low throughput, only one time-point can be analyzed for one mouse litter, possible position effect from the random construct insertion site, limitations caused by possible, and enhancer-promoter specificity. Furthermore, detection of silencer elements would require development of a special experimental system.

Some problems of mouse transient transgenics could be overcome by use of zebra fish or frog embryos with green fluorescent protein to allow visualization through development. This is a potential solution assuming that the development is sufficiently conserved for the analysis to apply in mammals.

Large genomic constructs

Large genomic constructs in bacterial or yeast artificial chromosomes (BACs or YACs) can be used to produce transient transgenic mouse embryos (Spitz et al., 2003). The gene of interest is tagged with reporter gene, and the expression driven from the large construct can be assayed. This approach has several advantages. First, reporter gene expression is driven by the endogenous promoter. Secondly, the positional effect of the insertion site is usually minimal with very large constructs. Third, manipulations of large construct (e.g. mutation of binding sites) allow identification of also silencers and insulators.

Mutating regulatory elements

An implicit proof of biological significance of an enhancer element is given by its targeted deletion from mouse genome. The examples of this approach are relatively few as the technique is time-consuming. Deletion of an enhancer element compromised T cell development (Mohrs et al., 2001). Dramatically, removal of approximately 1kb enhancer element of Sonic Hedgehog (Shh) located about 1 Mb away from its target gene causes truncation of mouse limbs (Sagai et al., 2005). In some cases targeted deletion of an enhancer element in mouse has not resulted in an apparent change in the phenotype (Ahituv et al., 2007). This latter result suggests that enhancers are likely to exhibit similar redundancy as protein-encoding regions of genes. There are most likely often many enhancers, which ensure that critical gene is expressed in the right place.

Chemical mutagenesis by N-ethyl-N-nitrosourea (ENU) has also been used in mice to induce point mutations, which may or may not, have an effect on enhancer function (Masuya et al., 2007). Point mutations, which alter mouse phenotype, indicate critical transcription factor binding sites in the enhancer element.

Chromatin-immunoprecipitation followed by parallel sequencing

None of the above mentioned methods of biological verification of mammalian enhancer elements are easily amenable to high-throughput format. A high-throughout approach for enhancer detection was used in a recent work. ChIP-seq method against transcriptional coactivator p300 done from embryonic limb, forebrain and midbrain ”fished out” enhancer

elements of these tissues (Visel et al., 2009). Even though the title of the work states that ”ChIP-seq accurately predicts tissue-specific activity of enhancer elements”, the enhancer elements were not predicted purely based on ChIP-seq data. The authors used also evolutionary conservation for the prediction of enhancer elements as they analyzed only elements, which were conserved in human and opossum. ChIP-seq is definitely a technique that allows genome-scale analysis of binding sites for a DNA-binding protein. However, this technique does not confirm that the identified sites have enhancer activity, i.e. that they can regulate transcription.

2.7 Single-nucleotide polymorphisms

Variations in a single nucleotide are the most common type of sequence variants within a species. Over 15 million single-nucleotide polymorphisms (SNPs) have been assigned to the human genome according to the Ensembl database (www.ensembl.org). SNPs are believed to be responsible for majority of the variation seen among individuals in human population.

2.7.1 Significance of SNPs

If SNPs are located in functional regions of the genome, they may cause variation in a population via different means. SNPs located in coding regions may result in functional consequences for example via missense or nonsense mutations, alterations in splicing, or mutation in genes coding RNAs. However, as majority of the genome is non-coding, many SNPs will fall in intergenic regions, and some of them will fall in regulatory elements.

During past decades, genetic analysis has been highly successful in finding disease genes and variants for many single gene disorders. The genetic dissection of common polygenic diseases has proven more challenging. The biggest value of SNP information is, at the moment, its usefulness in constructing haplotype maps that are used to find the genomic locations of causative variants for polygenic diseases. The international HapMap project has provided data that allows genome-wide association studies to identify susceptibility variants for human diseases (Frazer et al., 2007). Genome-wide association studies provide a means to dissect genetic determinants of common, multifactorial diseases. This work has high medical interest because common diseases affect large numbers of people.

Cancer is one example of a disease that can be studied by this approach. Common cancers have a genetic component in their etiology, and first

degree relatives have usually 2-4-fold risk of developing the same type of cancer (Goldgar et al., 1994). Genome-wide association studies have identified SNPs with susceptibility for common types of cancers, such as melanoma, breast, prostate, colorectal and lung cancer (Easton and Eeles, 2008). Susceptibility SNPs point genetic locations for targeted screening of causative variants. However, the molecular basis behind the associated cancer remains unknown for majority of susceptibility SNPs found so far. It is expected that the causative change lies somewhere in the vicinity of the susceptibility SNP, but its identification is often challenging.

2.7.2 Regulatory SNPs

Regulatory SNPs are SNPs that reside in regulatory elements and affect gene expression. It is not possible to predict the prevalence of regulatory SNPs due to the present lack of knowledge about regulatory elements (Hudson, 2003). Most of the identified regulatory SNPs are located in proximal promoters or introns, meaning within a few kilobases from the transcription start site (De Gobbi et al., 2006; Kim et al., 2009; Mangino et al., 2008; Munkhtulga et al., 2007). Some are located in distal enhancer elements (Rahimov et al., 2008; Steidl et al., 2007).

In the future years, our understanding of SNPs and their significance as source of variation in a population will certainly increase. Allelic variants of SNPs located in enhancers and other regulatory elements provide a potential source of variation in location, timing and intensity of gene expression. The variation caused by a single SNP might not be drastic, but the whole regulatory SNP complement of an individual could importantly shape the expression profile through time and space. Such differences in expression might explain variation in a multitude of traits, also size of organs and body parts, but more applicably susceptibility to diseases. In the future, susceptibility SNP analysis from patient's DNA sample could provide risk prediction and for example assessment of need for more invasive or expensive cancer screening procedures. Accumulation of comprehensive data about regulatory SNPs will take time. Understanding the functional relevance of regulatory SNPs will require extensive research into the logic of enhancers and other regulatory elements. The work in this thesis provides some of the first steps towards that goal.

3 AIMS OF THE STUDY

The motivataion for this work was to address the intriguing phenomenon of organ-specific growth. Our hypothesis was that growth *in vivo* is controlled by direct integration of growth factor-activated and tissue-specific signals on enhancer elements of critical cell cycle regulatory genes, such as the *Myc* genes (Figure 5). An enhancer element could drive expression of a critical growth gene in a specific tissue, and this mechanism could explain how specific tissues grow at different times and intensities.

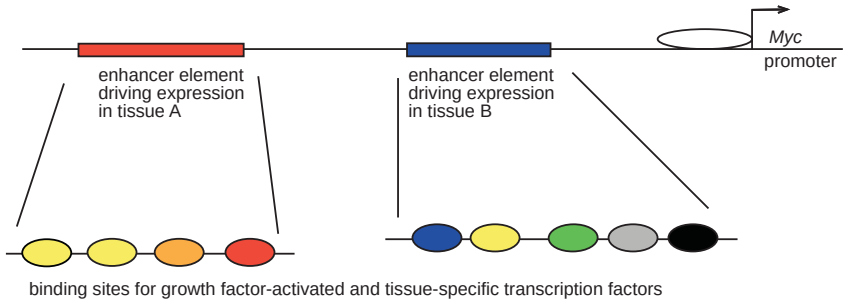


Figure 5. Hypothesis. Different enhancer elements drive expression of the target gene in different tissues at specific developmental time points.

The specific aims of this study were:

1. To determine binding specificities of transcription factors which are involved in growth control.
2. To find evolutionarily conserved enhancer elements, which drive expression in an organ-specific manner in the vicinity of genes regulating cell cycle progression.

4 MATERIALS AND METHODS

The methods used in the original articles of this thesis are summarised in the following table. The number indicates in which publication the method has been used (see list of original publications on page 13).

Method	Publication
Cell culture and transfection	I, II, III
Chromatin immunoprecipitation	III
Collection of samples from cancer patients and controls	III
Electrophoretic mobility shift assay	II, III
Exon array data analysis	III
Generating transgenic mouse embryos and mouse lines	II, III
Histological analysis	II
In silico sequence analysis	II, III
In situ hybridization	II, III
LacZ staining of tissues	II, III
Microarray gene expression experiments	III
Mouse breeding	II, III
PCR genotyping	II, III
Protein-DNA binding assay	I, II, III
RNA extraction and real-time PCR	III
Recombinant DNA techniques	I, II, III
Recombinant protein production in mammalian cells	I, II, III
SNP array experiments and data analysis	III

5 RESULTS AND DISCUSSION

5.1 High-throughput assay for determining specificity and affinity of protein-DNA binding interactions

Lack of knowledge on the binding specificities of transcription factors hinders our ability to predict and study transcriptional regulatory elements in genomic DNA. This is currently the main obstacle in advancing our understanding of code of transcriptional regulation. The binding specificities of transcription factors are the basis, the ”letters” of this code. Once we know the complete ”alphabet”, it will be much easier to try and read the language of transcriptional regulation of gene expression. Binding specificity is a crucial aspect of the function of any DNA-binding protein.

We have developed a method for high-throughput measurement of protein-DNA binding interactions (I: Figure 1, overview of the method). Briefly, the method requires prior knowledge of a high-affinity binding sequence (consensus). The DNA-binding protein to be studied is cloned in fusion with the *Renilla* luciferase enzyme. The fusion protein is then incubated with double-stranded biotinylated consensus DNA oligonucleotides and double-stranded non-biotinylated competitor DNA oligonucleotides. The amount of fusion protein that coprecipitates with the biotinylated oligonucleotide in a streptavidin-coated microtiter plate is measured by a luminometer. If the DNA-binding protein binds to the competitor oligonucleotide, the luminescence will be reduced compared to competition with a negative control DNA oligonucleotide (e.g. with bases of the binding site in a random order). The level of competition measures the affinity of DNA-binding protein to the competitor DNA.

We used this method to determine the binding specificity of several transcription factors: GLIs1-3, Ci, Tcf4 and c-Ets1 (II: Figure 1D). The results show that the method can be applied for analysis of transcription factors from different DNA-binding protein families, including zinc-finger (GLI), high-mobility-group (Tcf4) and ETS families (c-Ets1).

Various methods for determining DNA-binding specificities of transcription factors are available (see section 2.5.1 in review of literature). The method described here is suitable for high-throughput analysis and automation in microwell format. The method measures the inherent affinity of a transcription factor to DNA (rather than occupied sites), so it yields

high quality weight matrices for location of binding sites in genomic DNA sequence by computational analysis. This method is relatively inexpensive and does not require very specialized reagents or technology in the laboratory. Limitation of our method is the requirement of prior knowledge of a high-affinity binding site. However, often some high affinity site is known for a transcription factor, or it can be determined, for example by ChIP or SELEX. In comparison to microarray technology, the microwell-based protein-DNA assay described here gives a smaller number (typically 96 or 385) of more accurate measurements from a larger number of samples. However, analysis of every permutation of n base pair binding site is more straight-forward in microarray format than by our assay (Badis et al., 2008; Berger et al., 2006; Warren et al., 2006; Zhu et al., 2009).

This assay of specificity and affinity of protein-DNA binding interactions can also be used to address other experimental questions besides transcription factor binding specificity for DNA. Additional applications for this assay include analysis of effects of post-translational modifications, mutagenesis, and small-molecule, protein or antibody libraries on protein-DNA binding.

5.2 Enhancer Element locator

Prediction of enhancer elements is challenging because of the size of mammalian genomes and the lack of precise knowledge of the ”rules” defining enhancer element structure and conservation. Enhancer elements are probably a highly variable entity both in size and sequence composition. We do know that enhancer elements have transcription factor binding sites in them, but the fact that transcription factor binding motifs are short and degenerate further complicates the task. Indeed, individual binding sites can be found all over genomes, and the majority of them are not functionally significant.

To search for enhancer elements, we have developed a computer program called enhancer element locator (EEL), that predicts enhancer elements (Palin, 2007; Palin et al., 2006). The enhancer model of EEL is based on physical interactions that occur within an enhancer element. EEL searches for conservation of enhancer elements in two evolutionarily related species. Functionality of an enhancer element is transmitted via the protein complex that binds to it. Following this logic, the capacity to bind the same arrangement of transcription factors is searched by EEL as indication of conservation of an enhancer in two species. Here, ‘arrangement’ means

binding sites of specific transcription factors and their distances from each other on genomic DNA sequence. Therefore, conservation of arrangement does not mean conservation of plain DNA sequence. As input, EEL takes two orthologous DNA sequences and positional weight matrices describing the binding specificities of a suitable set of transcription factors. EEL locates the binding sites in the provided orthologous DNA sequences and then aligns the binding sites to look for conserved arrangements of sites. EEL gives a score for each predicted enhancer element. The score integrates values for three components: affinity, clustering and conservation of binding sites (II: Figure 2A). The real occupancy of a transcription factor binding site on DNA depends, in addition to the protein's affinity for DNA, on secondary interactions between adjacent transcription factors, as well as transcription factors and other proteins in the complex. Currently, there is not enough data available to model all the secondary interactions taking place in a protein complex on an enhancer element. The most detailed structural model of protein complex on an enhancer has been made on the interferon-β enhanceosome (Panne et al., 2007). This is a single case where the interactions between transcription factors and their contacts to DNA have been described in a highly conserved enhancer element. More data is needed for more complete understanding of the secondary interactions. In the EEL scoring scheme there are two elements, which reflect the possible secondary interactions: distance between adjacent binding sites within an enhancer (clustering), and conservation of distances between sites in orthologous enhancers (II: Figure 2A) (Palin et al., 2006); (Palin, 2007). There is evidence that single-nucleotide substitutions, small insertions, and small deletions occur freely within regulatory sequences, but that large insertions and deletions (>20 base pairs) are statistically almost absent within the regulatory sequences, which is consistent with the logic of EEL model (Cameron et al., 2005).

The EEL program was designed to predict mammalian enhancers. However, the relative weights of the components (affinity, clustering and conservation) are adjustable by changing values for λ, μ, ν and ξ (II: Figure 2A). Therefore the program can be adapted to search for enhancers, which characteristics differ from mammals. As distal enhancer elements may be located very far from the coding sequences, it was important that EEL program can efficiently analyze long DNA sequences: hundreds of kilobases of DNA sequence using dozens of binding motifs are analyzed in minutes. The most important characteristic of EEL, in comparison to other enhancer prediction methods available, is that it is based on a model of the physical interactions within an enhancer element (examples of other

enhancer prediction software in section 2.6.4. of review of literature). EEL is built to reflect the protein assembly allowing enhancer function and ultimately transcriptional regulation of gene expression. Even though our knowledge of protein-protein interactions in the process is limited, the EEL program can hopefully be developed further in future, as we reach more detailed understanding of the mechanisms behind function of enhancer elements.

Limitations of EEL relate to its intrinsic properties. For an enhancer element to be located by EEL, the enhancer needs to be composed of a conserved clustering of known transcription factor binding sites, which binding specificities are known. Should there exist an enhancer element, which had very few transcription factor binding sites, or in which the binding sites were very sparsely spread out, the chances of finding those enhancer elements by EEL are poor. Also, suitable evolutionary distance is required between the orthologous sequences analyzed. If the species are too closely related, such as mouse and rat, or human and chimpanzee, the sequence flanking the transcription factor binding sites has not diversified enough reveal the enhancers; the sequence is conserved and EEL predicts enhancers everywhere. As binding motifs are short and degenerate, they can be found in any sufficiently long DNA sequence, and thus EEL predicts enhancers in any conserved sequence, such as coding regions, repetitive sequences, and other evolutionarily highly conserved regions (e.g. HOX gene cluster in mammals). Therefore, the interpretation of EEL results requires understanding of the sequence architecture of the region that is analyzed. At the other end of the conservation spectrum, there are limitations in finding enhancers in genomic sequence from too distantly related species. If the transcription factor binding-sites, or their order, are not conserved, it is not possible to locate enhancer elements by EEL. For example in *Drosophila* species, there is considerable variation of binding site architecture of regulatory modules even when their function is conserved (Simpson and Ayyar, 2008). This might relate to the divergent evolution between *Drosophila* species that are actually more distantly related than different mammalian species.

In general, selection of species, which genomic DNA is analyzed, is very important in the prediction of enhancer elements using comparative genomic approaches. If a specific trait is studied, the species should have phenotypic similarity in that trait. More distantly related species have less similarity in enhancer elements. Species are expected to also have non-conserved enhancers that regulate species-specific patterns of gene

expression. Sometimes an enhancer element may be conserved but the pattern of gene expression is not. Inclusion of multiple species in the analysis generally increases the specificity but decreases the sensitivity in the predictions.

5.3 Genome-wide prediction of mammalian enhancer elements

We performed a genome-wide prediction of enhancer elements by analyzing 20,173 orthologous human-mouse gene pairs by EEL (II). The analyzed genomic sequence contained coding regions and 100 kb flanking sequence in both directions for each gene. The results were stored in a relational database, from which data about the binding sites, enhancers, and genes could be extracted and further analyzed (II: Figure 3A).

To validate the genome-wide prediction of enhancer elements, we extracted out different sets of data from the relational database, made predictions and tested whether they were correct. The approaches used in the validation were analysis of overrepresentation of a TF in a similarly expressed set of genes (II: Figure 3B, C), *in silico* prediction target genes of signaling pathways (II: Figures 4 and 5), and application of EEL analysis to a biological problem of organ-specific growth control (II: Figures 6 and 7).

The experimental validation of genome-scale enhancer prediction in mammals is currently challenging. Tens of thousands of enhancer elements are predicted by computational analysis, but the biological validation of even a dozen mammalian enhancers in transgenic mouse embryos is laborious. Moreover, the experimental validation of regulatory elements is always context dependent. Analysis of mammalian embryos requires removal of the embryos from the uterus. Therefore only one time-point can be studied in one experiment, and a negative result does not rule out the possibility, that the element would be functional in a different time point and/or cell type. Also, the very same element could have critical function in many tissues and stages, or it may act as a silencer in another context. Even more time-consuming is validation of enhancer elements by targeted deletion in mouse. Possible redundancy of enhancer elements can further increase the challenge, so that all the redundant enhancer elements need to be targeted before potential phenotype becomes apparent.

In this work, we obtained a general picture of what a functional mammalian enhancer element looks like. It is important to bear in mind, however, that

Figure 6. EEL alignment of the enhancer element located in the second intron of *N-Myc*. The enhancer drives expression in the developing tooth. Transcription factor binding sites indicated by yellow boxes and respective TFs in green.

choice of the method, in this case the EEL software, produces its own biases. Most notably, EEL is not likely to find enhancer elements, which would be very different from what is assumed by the EEL model. Nevertheless, the results show that in the genome there seems to be thousands of enhancer elements, which approximately fulfill the characteristics outlined above. A typical mammalian enhancer element based on this work is about 1 kb long and contains about a dozen transcription factor binding-sites (example in Figure 6). This work did not aim at intensely scrutinizing the borders of enhancer elements. Excluding predicted enhancer elements which have alternative causes for sequence conservation, such as coding region or repetitive sequence, an high-scoring EEL-predicted enhancer has a good potential to be functional. To this end, knowledge of signaling pathway that regulates expression of predicted target gene improves the odds of finding a biologically relevant enhancer element. If certain signal pathway is known to regulate the gene, predicted enhancer elements, which contain binding sites for the transcription factor that is activated by this pathway and also have a high EEL score, are prominent candidates for functional enhancer elements.

5.4 Tissue-specific enhancers of *c-Myc* and *N-Myc* loci may drive organ-specific growth

To address the biological problem of organ-specific growth control, we chose to study two *Myc* genes, *c-Myc* and *N-Myc*, in more detail. We analyzed 200 kb genomic sequence around coding region of these genes and were able to identify several enhancer elements that drive expression of a marker gene in a tissue-specific manner (II: Figures 6 and 7).

Two enhancer elements of the *N-Myc* locus are particularly interesting in the light of data available from the tissues where they are expressed (II: Figure 7B, C). Enhancer element or *cis*-module 7 (CM7), located in the second intron, drives expression in the maxillary arch derivatives, including the developing tooth bud (II: Figure 7B, E). This enhancer contains, among other transcription factor binding sites, one conserved binding site for GLI, the transcription factor that is activated by the Hedgehog pathway (Figure 6). At E12.5 the enhancer drives expression in the dental epithelium, coinciding with Shh expression (Bitgood and McMahon, 1995). Shh has been reported to act as a mitogen and survival factor in the tooth development (Cobourne et al., 2001), and our results suggest that this effect is mediated via N-Myc protein. In addition to the signal pathway-activaed

transcription factor binding site (GLI), this enhancer contains a dozen other sites. Of those, at least Ahr-ARNT, SP1, and Smad5 are expressed in the tooth area and represent potential tissue-specific transcription factors of this enhancer element (Figure 6) (Aitola and Pelto-Huikko, 2003; Bouwman et al., 2000; Murashima-Suginami et al., 2008).

The second enhancer element of *N-Myc*, CM5, is located 65 kb downstream of the *N-Myc* transcription start site. It contains a conserved GLI binding site. At E 12.5, this enhancer drives expression in the forebrain and in the dorsal aspect of the neural tube (II: Figure 7C, F, G). Furthermore, in newborn mice at postnatal day 3 (PN3) the expression is driven in the cerebellar granule neuron progenitors (CGNPs) of the external granule cell layer of the cerebellum (II: Figure 7H). It has been shown in cell culture and mouse experiments that N-Myc acts downstream of Shh signaling during CGNP proliferation, and that N-Myc is essential downstream effector of Shh signaling during cerebellar growth (Hatton et al., 2006; Kenney et al., 2003; Oliver et al., 2003).

Multiple tissue-specific enhancers for a single gene, as presented here for *N-Myc*, provide explanation for the mechanism how growth can be controlled in different ways in distinct tissues and organs. Also, we located two enhancer elements, which both contain conserved GLI-binding sites and are potentially Hh pathway responsive, but drive expression in different tissues. It is likely that the activity of the enhancer element is controlled, in addition to Hh pathway, by transcription factors which expression is restricted to specific tissue(s).

Methods to further validate the importance of these enhancer elements in regulation of *N-Myc* expression include 3C techniques to show physical contact between the enhancer and promoter of *N-Myc*, and targeted deletion of these enhancer elements in mouse to show possible phenotypic effect.

5.5 Regulatory SNP located in an enhancer element affects cancer susceptibility

The human SNP rs6983267, located in the long arm of chromosome 8, carries either T or G allele. The G allele has been identified to be associated with an increased susceptibility to colorectal cancer in several genome-wide association studies (Haiman et al., 2007; Tomlinson et al., 2007; Zanke et al., 2007). The large genomic region carrying the SNP rs6983267 is amplified in colorectal cancers (III: Figure 1A, B). We found that the SNP

rs6983267 falls in a Tcf4 binding site, which is located in a high EEL-scoring enhancer element (III: Figure 2A). The G allele the final position of the Tcf4 binding site results in 1.6 fold increase in the affinity of this binding site compared to a Tcf4 binding site with a T allele in this position (III: Figure 2B). In genome-wide ChIP assay, this Tcf4 site gives the highest signal within 1 Mb of *c-Myc* (III: Figure 3D). Tcf4 preferably binds to the binding site with the risk-allele G in colorectal cancer cell lines (III: Figure 3F, G). In cell culture where the Wnt pathway is active, the enhancer containing the risk-allele G drives 1.5-fold higher expression of a marker gene than the enhancer with the T allele (III: Figure 2D). In transient transgenic mouse embryos, the predicted enhancer element drives expression of a marker gene in a pattern that is consistent with regulation by Wnt pathway (III: Figure 4). These and other lines of evidence in III indicate that predisposition to colorectal cancer linked to the G allele in rs6983267 increases Tcf4 binding affinity and thus enhances responsiveness to Wnt signaling.

The gene closest to this enhancer is *c-Myc*, located 335 kb from the enhancer. We propose that *c-Myc* is the target gene of the enhancer element that contains the SNP rs6983267. The distance to the proposed target gene is long in comparison to previously published regulatory SNPs in other genes, maximum published distance being 16 kb (Steidl et al., 2007). But it is well within the range of known functional enhancer elements, as a known functional enhancer element of *Shh* is 1 Mb from its target (Sagai et al., 2005). Together with the Wnt pathway, c-Myc has an important role in colorectal cancer. The Wnt pathway is activated in 90% of colorectal cancers, and *c-Myc* has been indentified as a target of the Wnt pathway (Bienz and Clevers, 2000; He et al., 1998). c-Myc expression is required for the tumorigenic phenotype resulting from *APC* loss in mice (Sansom et al., 2007). Our work provides a plausible mechanistic explanation for the link between the risk-allele G in the SNP rs6983267 and colorectal cancer.

It is challenging to explicitly show a correlation between rs6983267 genotype and the expression level of c-Myc *in vivo*. We have attempted to show the correlation in Epstein-Barr virus transformed lymhoblastoid cell lines of HapMap individuals and in colorectal cancer tumor samples. Neither of these is optimal starting materials for the analysis. The function of enhancer elements is tissue-specific; it is likely that the intestinal enhancer element is not functional in other tissues. There is indeed evidence that Wnt pathway is not active in Epstein-Barr virus transformed lymhoblastoid cell lines (Everly et al., 2004). Colorectal tumors are better

starting material in the sense that they originate from the tissue where the enhancer is expected to be active. However, colorectal tumors are heterogeneous cell populations; they contain various genetic alterations and differ in their Wnt signaling status (Fodde and Brabletz, 2007). Therefore, the correlation between the rs6983267 genotype and c-Myc expression level in tumor tissue may well be perturbed, even if it originally was critical in tumor development. Ideally the correlation between rs6983267 genotype and the expression level of c-Myc would be analyzed in normal colon. However, obtaining sufficient material for RNA extraction from the few stem cells at the bottoms of the intestinal crypts, where Wnt signaling is active, remains a challenge. One possibility would be to try and show the correlation in mice, as obtaining fresh tissue samples from mice is easier.

The risk allele G is very common in human population, with 50% allele frequency in Caucasians and almost 100% frequency in people of African origin (Haiman et al., 2007). In fact, the G allele is most likely the ancestral form. Consistent with high prevalence of the risk allele, African Americans are more likely to develop colorectal cancer and die of it than Caucasians in the United States (Polite et al., 2006). Even though there could be environmental reasons to account for this, the younger age at presentation and the high mortality rates among the younger cohorts of African American patients suggest a genetic difference between the two groups (Polite et al., 2006).

All the cancer susceptibility loci found in genome-wide association studies confer a relatively low risk of cancer, less than 1.5-fold (Haiman et al., 2007). Homozygosity for the G allele of SNP rs6983267 increases colorectal cancer risk ~1.5 fold. It is considered unlikely that loci with stronger effects will be found in the future, considering the extensive size and the coverage of the genome-wide association studies conducted so far (Easton and Eeles, 2008). Likely, this is the nature of susceptibility loci for common, low-penetrance, multifactorial diseases. Susceptibility loci will be numerous and the effect of each is expected to be minor. Within the population the diseases are common compared to single gene disorders, for example 5% of population in Western world develops colorectal malignancies (Bienz and Clevers, 2000). But the penetrance is low, and the prevalence of a risk allele is very high in the population.

Due to its importance in development of colorectal cancer, the Wnt pathway is a self-evident target for rational cancer therapies. Our results suggest that the Wnt pathway also has potential as a target for personalized cancer prevention strategies.

5.6 Tissue-specific enhancers may explain tumor-type specificity of oncogenes

Cancers of specific organs exhibit typical genetic alterations. Where does this specificity arise?

One explanation for this phenomenon is the tissue-type specific regulation of *Myc* genes (or other important oncogenes and tumor suppressor genes). The majority of human malignancies express one or more *Myc* genes (Pelengaris et al., 2002). Enhancer elements composed of many transcription factor binding-sites integrate signals from ubiquitously expressed and tissue-specific transcription factors as well as signaling pathway-activated transcription factors. Mutations affecting specific signaling pathway could lead to unrestricted growth only in tissues where additional factors allow the activation of *Myc* gene expression via a tissue-specific enhancer element. This could explain why certain genetic alterations cause cancer only in particular tissues or organs.

In III, we propose an enhancer element, which could transmit tissue-specificity of a constitutively activated Wnt signaling pathway in causing colorectal cancer. Medulloblastoma, the most common childhood malignancy of the central nervous system, is dependent on Shh signaling via N-Myc for unrestricted proliferation of CGNPs (Zindy et al., 2006). The enhancer element CM7, described in II, could be a candidate to pass the signal of activated Hh pathway to N-myc in formation of medulloblastoma.

6 CONCLUSIONS

This work embarked to address the phenomenon of organ-specific growth control. Our hypothesis was that growth *in vivo* is controlled by direct integration of growth factor-activated and tissue-specific signals on enhancer elements of critical cell cycle regulatory genes, such as the *Myc* genes The hypothesis focused the area of interest into transcriptional regulation via enhancer elements and how organ-specific growth control could function via these elements.

The first obstacle in this work was that only limited information exists about the DNA-binding specificities of even well known transcription factors. Therefore, we set out to develop a high-throughput method for measuring transcription factor DNA-binding specificities. This method is applicable to a range of transcription factors, as it was used to determine binding specificities of three different families of DNA-binding proteins: zinc finger domain (GLI), HMG box (Tcf4) and ETS domain (Ets-1). This high-throughput assay is generally applicable method for addressing biological problems related to protein-DNA binding interactions.

The existing computational resources were not sufficient to efficiently locate distal mammalian enhancer elements up to hundreds of kilobases away from the coding regions of their target genes. To overcome this limitation, we developed EEL, a computational tool by which enhancer elements can be predicted. We used this program to analyze several vertebrate genomes. We performed a genome-wide prediction of human and mouse enhancer elements, which provides a large database for different uses. For example, we showed that the genome-wide data is useful in identification of Hh and Wnt target genes. It could also be used for prediction of target genes of other signaling pathways.

We found multiple tissue-specific enhancers in mouse *c-Myc* and *N-Myc* genes, thereby uncovering a potential mechanism by which growth and cancer can occur in organ-specific fashion.

Moreover, we have located a regulatory SNP in a Wnt signaling-responsive enhancer element 335 kb upstream of *c-Myc*. This part of the work implicates a mechanistic explanation for the link between this SNP and occurrence of colorectal cancer. Genome-wide association studies are identifying increasing number of SNPs associated with common diseases, but usually the causative mechanism remains obscure. This work provides

an examle on how common, low-penetrance disease SNP variant could transmit the susceptibility to diseases, such as cancer.

The future work in this project will involve generating mice with targeted conditional deletion of the tissue-specific enhancer elements of *N-Myc* gene, which drive expression in the tooth and cerebellum, and the enhancer element of *c-Myc*, which drives the Wnt-responsive expression pattern and contains the SNP rs6983267 associated with colorectal cancer. Deletion of an enhancer element in mouse will provide an organism level evidence of its significance in growth of the respective organs. As the enhancer element could be functional in multiple tissues and developmental stages, the targeting is best done conditionally. Thus we can delete the enhancer in a specific tissue and study the effect there.

It is clear that not all the regulatory elements of the *c-Myc* and *N-Myc* genes have yet been discovered. Future work will uncover more enhancer elements and reveal more complete picture of the transcriptional regulation of *c-Myc* and *N-Myc*.

Even though the *Myc* genes are critical to growth in many situations, there are also other genes which transcriptional regulation would be interesting to study in respect to organ-specific growth control. After the beginning of this thesis work, cdk1 and cyclins A2 and B1 have been shown to be very critical to growth *in vivo*, as their deletion results in very early embryonic lethality (unlike other cdks and cyclins). It could be interesting to investigate the enhancer elements of these genes. Also multiple lines of evidence have emerged indicating that the *IGF* genes are critical to growth control. Dissection of their regulatory elements could reveal a very interesting growth regulatory potential.

7 ACKNOWLEDGEMENTS

This work was conducted in the laboratory of professor Jussi Taipale in Biomedicum Helsinki, at University of Helsinki and National Public, now National Institute for Health and Welfare, during the years 2003-2009.

I want to express my deepest gratitude to my supervisor, Jussi Taipale. Without you none of this would have happened. Thank you for taking me onboard this great journey.

I am grateful to the Biomedicum fifth floor infrastructure, Cancer biology program and Genome-scale biology program, and their directors professor Kari Alitalo and professor Tomi Mäkelä. Also, I thank Department of Molecular Medicine and its head adjunct professor Anu Jalanko for the facilities in the third floor.

Professors Tapio Palva and Minna Nyström in division of genetics, University of Helsinki, as well as Jonna Katajisto and Marjo Kestilä, gave crucial help with various official steps of attaining PhD degree.

The reviewers, professor Jukka Jernvall and professor Jorma Palvimo, are acknowledged for the insightful comments on the thesis.

I have been very fortunately to have great collaborators: Kimmo Palin, professor Esko Ukkonen, Natalia Sinjushina, and professor Juha Partanen. Kimmo, your understanding of biology goes so much further than my understanding of computer science ever will. Natalia, I fondly remember the nice moments in the lab with you. I thank Professor Lauri Aaltonen and all the collaborators in the SNP paper for the opportunity to be part of exciting and important research. It was comforting to share the last-minute PhD stress and discussions with Sari Tuupanen.

I thank everybody in the Taipale lab for being there in the ups and downs. Markku, thanks for all the years in the lab; I hope a tiny bit of your experimental skill diffused into me during those years. In relation to this thesis work, I especially want to acknowledge Minna Taipale, Jussi Taipale, Mikael Björklund, Mia Mönkkönen, Mikko Turunen, Arttu Jolma, Anna Saramäki, and Martin Bonke for reading my thesis and giving valuable comments – I would have never finished without you telling me that it is good already! Thanks go to Maria Sokolova, Song-Ping Li, Gong-Hong Wei, Lin Feng, Jian Yan, Teemu Kivioja, Fu-Ping Zhang, James Thompson, and also past members of Taipale lab. Sini Miettinen, Ritva

Nurmi, and Tiia Pelkonen have helped me so many times with various practical matters.

I thank everyone in the animal facility for important work. Essi Kaiharju and Pirjo Ranta gave great help in doing a lot of genotyping. Kirsi Salonen's and Riikka Pakarinen's skillful work in transgenic unit was vital for this project.

There are so many friends and colleagues in Biomedicum that it is impossible to mention you all, for example people in Mäkelä lab, Ojala lab, Laiho lab, Keski-Oja lab, and many others. Thank you for the help and discussions. Also friends and teachers in Viikki campus are warmly acknowledged. Professor Jim Schröder first opened the door to science for me. Professor Irma Thesleff was the first to believe in this project - already before the work started.

My dear friends outside science: Elliot, Merja, Sulo, Jade, Jussi, Suvi, Seela, Aleksis, Johanna, Sampo, Reetta, Tero, and Mira & co, thank you for being there along the way. During these years I have had so many important discussions with Paola, Daniela, Vivian, Saara, Mariana, Janice and all my friends abroad in both face to face and in emails. It has meant a lot to me.

My family has followed me struggle my way through the jungle of PhD work. Mum, Dad, Riikka, Ville, Elina, Iiro, Salma and Notte, thank you for your support.

Tero, I thank you for believing in me in the moments when I did not. Reko, you have showed me how to explore the world with fresh and keen eyes.

This work has been financially supported by the Academy of Finland, European Comission FP6 Regulatory Genomics Program, the Jenny and Antti Wihuri Foundation, the Magnus Ehrnrooth Foundation, the Finnish Cancer Organizations, the Ida Montin Foundation, the Ella and Georg Ehrnrooth Foundation, and the Maud Kuistila Memorial Foundation.

Outi Kallikas

Helsinki, May 2009

8 REFERENCES

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F.*, et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L. A., and Rubin, E. M. (2007). Deletion of ultraconserved elements yields viable mice. PLoS Biol *5*, e234.

Aitola, M. H., and Pelto-Huikko, M. T. (2003). Expression of Arnt and Arnt2 mRNA in developing murine tissues. J Histochem Cytochem *51*, 41-54.

Allan, M. F., Eisen, E. J., and Pomp, D. (2005). Genomic mapping of direct and correlated responses to long-term selection for rapid growth rate in mice. Genetics *170*, 1863-1877.

Badis, G., Chan, E. T., van Bakel, H., Pena-Castillo, L., Tillo, D., Tsui, K., Carlson, C. D., Gossett, A. J., Hasinoff, M. J., Warren, C. L.*, et al.* (2008). A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. Mol Cell *32*, 878-887.

Benos, P. V., Bulyk, M. L., and Stormo, G. D. (2002). Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res *30*, 4442-4451.

Berger, M. F., Badis, G., Gehrke, A. R., Talukder, S., Philippakis, A. A., Pena-Castillo, L., Alleyne, T. M., Mnaimneh, S., Botvinnik, O. B., Chan, E. T.*, et al.* (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell *133*, 1266-1276.

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., 3rd, and Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol *24*, 1429-1435.

Berns, K., Hijmans, E. M., Koh, E., Daley, G. Q., and Bernards, R. (2000). A genetic screen to identify genes that rescue the slow growth phenotype of c-myc null fibroblasts. Oncogene *19*, 3330-3334.

Berthet, C., Aleem, E., Coppola, V., Tessarollo, L., and Kaldis, P. (2003). Cdk2 knockout mice are viable. Curr Biol *13*, 1775-1785.

Bi, L., Okabe, I., Bernard, D. J., and Nussbaum, R. L. (2002). Early embryonic lethality in mice deficient in the p110beta catalytic subunit of PI 3-kinase. Mamm Genome *13*, 169-172.

Bi, L., Okabe, I., Bernard, D. J., Wynshaw-Boris, A., and Nussbaum, R. L. (1999). Proliferative defect and embryonic lethality in mice homozygous for a deletion in the p110alpha subunit of phosphoinositide 3-kinase. J Biol Chem *274*, 10963-10968.

Bienz, M., and Clevers, H. (2000). Linking colorectal cancer to Wnt signaling. Cell *103*, 311-320.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., *et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816.

Bitgood, M. J., and McMahon, A. P. (1995). Hedgehog and Bmp genes are coexpressed at many diverse sites of cell-cell interaction in the mouse embryo. Dev Biol *172*, 126-138.

Bouwman, P., Gollner, H., Elsasser, H. P., Eckhoff, G., Karis, A., Grosveld, F., Philipsen, S., and Suske, G. (2000). Transcription factor Sp3 is essential for post-natal survival and late tooth development. Embo J *19*, 655-661.

Brandeis, M., Rosewell, I., Carrington, M., Crompton, T., Jacobs, M. A., Kirk, J., Gannon, J., and Hunt, T. (1998). Cyclin B2-null mice develop normally and are fertile whereas cyclin B1-null mice die in utero. Proc Natl Acad Sci U S A *95*, 4344-4349.

Bushey, A. M., Dorman, E. R., and Corces, V. G. (2008). Chromatin insulators: regulatory mechanisms and epigenetic inheritance. Mol Cell *32*, 1-9.

Camargo, F. D., Gokhale, S., Johnnidis, J. B., Fu, D., Bell, G. W., Jaenisch, R., and Brummelkamp, T. R. (2007). YAP1 increases organ size and expands undifferentiated progenitor cells. Curr Biol *17*, 2054-2060.

Cameron, R. A., Chow, S. H., Berney, K., Chiu, T. Y., Yuan, Q. A., Kramer, A., Helguero, A., Ransick, A., Yun, M., and Davidson, E. H. (2005). An evolutionary constraint: strongly disfavored class of change in DNA sequence during divergence of cis-regulatory modules. Proc Natl Acad Sci U S A *102*, 11769-11774.

Carmo-Fonseca, M. (2002). The contribution of nuclear compartmentalization to gene regulation. Cell *108*, 513-521.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engstrom, P. G., Frith, M. C., *et al.* (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet *38*, 626-635.

Carroll, J. S., Liu, X. S., Brodsky, A. S., Li, W., Meyer, C. A., Szary, A. J., Eeckhoute, J., Shao, W., Hestermann, E. V., Geistlinger, T. R., *et al.* (2005). Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell *122*, 33-43.

Carroll, S. B. (2005). Evolution at two levels: on genes and form. PLoS Biol *3*, e245.

Charron, J., Malynn, B. A., Fisher, P., Stewart, V., Jeannotte, L., Goff, S. P., Robertson, E. J., and Alt, F. W. (1992). Embryonic lethality in mice homozygous for a targeted disruption of the N-myc gene. Genes Dev *6*, 2248-2257.

Chen, W. S., Xu, P. Z., Gottlob, K., Chen, M. L., Sokol, K., Shiyanova, T., Roninson, I., Weng, W., Suzuki, R., Tobe, K., *et al.* (2001). Growth retardation and increased apoptosis in mice with homozygous disruption of the Akt1 gene. Genes Dev *15*, 2203-2208.

Cheverud, J. M., Routman, E. J., Duarte, F. A., van Swinderen, B., Cothran, K., and Perel, C. (1996). Quantitative trait loci for murine growth. Genetics *142*, 1305-1319.

Cho, H., Thorvaldsen, J. L., Chu, Q., Feng, F., and Birnbaum, M. J. (2001). Akt1/PKBalpha is required for normal growth but dispensable for maintenance of glucose homeostasis in mice. J Biol Chem *276*, 38349-38352.

Cobourne, M. T., Hardcastle, Z., and Sharpe, P. T. (2001). Sonic hedgehog regulates epithelial proliferation and cell survival in the developing tooth germ. J Dent Res *80*, 1974-1979.

Conlon, I., and Raff, M. (1999). Size control in animal development. Cell *96*, 235-244.

Conlon, I., and Raff, M. (2003). Differences in the way a mammalian cell and yeast cells coordinate cell growth and cell-cycle progression. J Biol *2*, 7.

Cooper, S. (2004). Control and maintenance of mammalian cell size. BMC Cell Biol *5*, 35.

Crick, F. H. (1966). The genetic code. 3. Sci Am *215*, 55-60 passim.

Crickmore, M. A., and Mann, R. S. (2006). Hox control of organ size by regulation of morphogen production and mobility. Science *313*, 63-68.

Davis, A. C., Wims, M., Spotts, G. D., Hann, S. R., and Bradley, A. (1993). A null c-myc mutation causes lethality before 10.5 days of gestation in homozygotes and reduced fertility in heterozygous female mice. Genes Dev *7*, 671-682.

De Gobbi, M., Viprakasit, V., Hughes, J. R., Fisher, C., Buckle, V. J., Ayyub, H., Gibbons, R. J., Vernimmen, D., Yoshinaga, Y., de Jong, P.*, et al.* (2006). A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science *312*, 1215-1217.

Dekker, J. (2008). Gene regulation in the third dimension. Science *319*, 1793-1794.

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

Dong, J., Feldmann, G., Huang, J., Wu, S., Zhang, N., Comerford, S. A., Gayyed, M. F., Anders, R. A., Maitra, A., and Pan, D. (2007). Elucidation of a universal size-control mechanism in Drosophila and mammals. Cell *130*, 1120-1133.

Dubois, N. C., Adolphe, C., Ehninger, A., Wang, R. A., Robertson, E. J., and Trumpp, A. (2008). Placental rescue reveals a sole requirement for c-Myc in embryonic erythroblast survival and hematopoietic stem cell function. Development *135*, 2455-2465.

Dufresne, S. D., Bjorbaek, C., El-Haschimi, K., Zhao, Y., Aschenbach, W. G., Moller, D. E., and Goodyear, L. J. (2001). Altered extracellular signal-regulated kinase signaling and glycogen metabolism in skeletal muscle from p90 ribosomal S6 kinase 2 knockout mice. Mol Cell Biol *21*, 81-87.

Easton, D. F., and Eeles, R. A. (2008). Genome-wide association studies in cancer. Hum Mol Genet *17*, R109-115.

Echave, P., Conlon, I. J., and Lloyd, A. C. (2007). Cell size regulation in mammalian cells. Cell Cycle *6*, 218-224.

Edgar, B. A. (2006). From cell structure to transcription: Hippo forges a new path. Cell *124*, 267-273.

Everly, D. N., Jr., Kusano, S., and Raab-Traub, N. (2004). Accumulation of cytoplasmic beta-catenin and nuclear glycogen synthase kinase 3beta in Epstein-Barr virus-infected cells. J Virol *78*, 11648-11655.

Fantes, P., and Nurse, P. (1977). Control of cell size at division in fission yeast by a growth-modulated size control over nuclear division. Exp Cell Res *107*, 377-386.

Fausto, N., Campbell, J. S., and Riehle, K. J. (2006). Liver regeneration. Hepatology *43*, S45-53.

Fodde, R., and Brabletz, T. (2007). Wnt/beta-catenin signaling in cancer stemness and malignant behavior. Curr Opin Cell Biol *19*, 150-158.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., *et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

Freiman, R. N., Albright, S. R., Zheng, S., Sha, W. C., Hammer, R. E., and Tjian, R. (2001). Requirement of tissue-selective TBP-associated factor TAFII105 in ovarian development. Science *293*, 2084-2087.

Gabellini, D., Green, M. R., and Tupler, R. (2002). Inappropriate gene activation in FSHD: a repressor complex binds a chromosomal repeat deleted in dystrophic muscle. Cell *110*, 339-348.

Gerhart, J., and Kirschner, M. (2007). The theory of facilitated variation. Proc Natl Acad Sci U S A *104 Suppl 1*, 8582-8589.

Glass, C. K., and Rosenfeld, M. G. (2000). The coregulator exchange in transcriptional functions of nuclear receptors. Genes Dev *14*, 121-141.

Goldgar, D. E., Easton, D. F., Cannon-Albright, L. A., and Skolnick, M. H. (1994). Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. J Natl Cancer Inst *86*, 1600-1608.

Gregory, T. R. (2004). Mammal erythrocyte sizes; http://www.genomesize.com/cellsize/mammals.htm.

Haiman, C. A., Le Marchand, L., Yamamato, J., Stram, D. O., Sheng, X., Kolonel, L. N., Wu, A. H., Reich, D., and Henderson, B. E. (2007). A common genetic risk factor for colorectal and prostate cancer. Nat Genet *39*, 954-956.

Hatton, B. A., Knoepfler, P. S., Kenney, A. M., Rowitch, D. H., de Alboran, I. M., Olson, J. M., and Eisenman, R. N. (2006). N-myc is an essential downstream effector of

Shh signaling during both normal and neoplastic cerebellar growth. Cancer Res *66*, 8655-8661.

Hatton, K. S., Mahon, K., Chin, L., Chiu, F. C., Lee, H. W., Peng, D., Morgenbesser, S. D., Horner, J., and DePinho, R. A. (1996). Expression and activity of L-Myc in normal mouse development. Mol Cell Biol *16*, 1794-1804.

He, T. C., Sparks, A. B., Rago, C., Hermeking, H., Zawel, L., da Costa, L. T., Morin, P. J., Vogelstein, B., and Kinzler, K. W. (1998). Identification of c-MYC as a target of the APC pathway. Science *281*, 1509-1512.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W.*, et al.* (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature.

Henery, C. C., Bard, J. B., and Kaufman, M. H. (1992). Tetraploidy in mice, embryonic cell number, and the grain of the developmental map. Dev Biol *152*, 233-241.

Hobert, O. (2008). Gene regulation by transcription factors and microRNAs. Science *319*, 1785-1786.

Hudson, T. J. (2003). Wanted: regulatory SNPs. Nat Genet *33*, 439-440.

Jacobsson, L., Park, H. B., Wahlberg, P., Fredriksson, R., Perez-Enciso, M., Siegel, P. B., and Andersson, L. (2005). Many QTLs with minor additive effects are associated with a large difference in growth between two selection lines in chickens. Genet Res *86*, 115-125.

Kenney, A. M., Cole, M. D., and Rowitch, D. H. (2003). Nmyc upregulation by sonic hedgehog signaling promotes proliferation in developing cerebellar granule neuron precursors. Development *130*, 15-28.

Kenney-Hunt, J. P., Vaughn, T. T., Pletscher, L. S., Peripato, A., Routman, E., Cothran, K., Durand, D., Norgard, E., Perel, C., and Cheverud, J. M. (2006). Quantitative trait loci for body size components in mice. Mamm Genome *17*, 526-537.

Kim, K., Sung, Y. K., Kang, C. P., Choi, C. B., Kang, C., and Bae, S. C. (2009). A regulatory SNP at position -899 in CDKN1A is associated with systemic lupus erythematosus and lupus nephritis. Genes Immun.

King, M. C., and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. Science *188*, 107-116.

Kiyokawa, H., Kineman, R. D., Manova-Todorova, K. O., Soares, V. C., Hoffman, E. S., Ono, M., Khanam, D., Hayday, A. C., Frohman, L. A., and Koff, A. (1996). Enhanced growth of mice lacking the cyclin-dependent kinase inhibitor function of p27(Kip1). Cell *85*, 721-732.

Kloster, M., Tang, C., and Wingreen, N. S. (2005). Finding regulatory modules through large-scale gene-expression data analysis. Bioinformatics *21*, 1172-1179.

Knobloch, J., Shaughnessy, J. D., Jr., and Ruther, U. (2007). Thalidomide induces limb deformities by perturbing the Bmp/Dkk1/Wnt signaling pathway. Faseb J *21*, 1410-1421.

Knoepfler, P. S., Cheng, P. F., and Eisenman, R. N. (2002). N-myc is essential during neurogenesis for the rapid expansion of progenitor cell populations and the inhibition of neuronal differentiation. Genes Dev *16*, 2699-2712.

Knuppel, R., Dietze, P., Lehnberg, W., Frech, K., and Wingender, E. (1994). TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. J Comput Biol *1*, 191-198.

Kolbe, D., Taylor, J., Elnitski, L., Eswara, P., Li, J., Miller, W., Hardison, R., and Chiaromonte, F. (2004). Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. Genome Res *14*, 700-707.

Kosak, S. T., and Groudine, M. (2004). Gene order and dynamic domains. Science *306*, 644-647.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W*., et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lawrence, P. A., Struhl, G., and Casal, J. (2008). Do the protocadherins Fat and Dachsous link up to determine both planar cell polarity and the dimensions of organs? Nat Cell Biol *10*, 1379-1382.

Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet *12*, 1725-1735.

Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. Nature *424*, 147-151.

Li, B., Carey, M., and Workman, J. L. (2007). The role of chromatin during transcription. Cell *128*, 707-719.

Lobanenkov, V. V., Nicolas, R. H., Adler, V. V., Paterson, H., Klenova, E. M., Polotskaja, A. V., and Goodwin, G. H. (1990). A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. Oncogene *5*, 1743-1753.

Majerus, M. E., and Mundy, N. I. (2003). Mammalian melanism: natural selection in black and white. Trends Genet *19*, 585-588.

Malik, S., and Roeder, R. G. (2000). Transcriptional regulation through Mediator-like coactivators in yeast and metazoan cells. Trends Biochem Sci *25*, 277-283.

Malumbres, M., Sotillo, R., Santamaria, D., Galan, J., Cerezo, A., Ortega, S., Dubus, P., and Barbacid, M. (2004). Mammalian cells cycle without the D-type cyclin-dependent kinases Cdk4 and Cdk6. Cell *118*, 493-504.

Mangino, M., Brouilette, S., Braund, P., Tirmizi, N., Vasa-Nicotera, M., Thompson, J. R., and Samani, N. J. (2008). A regulatory SNP of the BICD1 gene contributes to telomere length variation in humans. Hum Mol Genet *17*, 2518-2523.

Martins, R. A., Zindy, F., Donovan, S., Zhang, J., Pounds, S., Wey, A., Knoepfler, P. S., Eisenman, R. N., Roussel, M. F., and Dyer, M. A. (2008). N-myc coordinates retinal growth with eye size during mouse development. Genes Dev *22*, 179-193.

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet *7*, 29-59.

Masuya, H., Sezutsu, H., Sakuraba, Y., Sagai, T., Hosoya, M., Kaneda, H., Miura, I., Kobayashi, K., Sumiyama, K., Shimizu, A.*, et al.* (2007). A series of ENU-induced single-base substitutions in a long-range cis-element altering Sonic hedgehog expression in the developing mouse limb bud. Genomics *89*, 207-214.

McGaughey, D. M., Vinton, R. M., Huynh, J., Al-Saif, A., Beer, M. A., and McCallion, A. S. (2008). Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. Genome Res *18*, 252-260.

McPherron, A. C., Lawler, A. M., and Lee, S. J. (1997). Regulation of skeletal muscle mass in mice by a new TGF-beta superfamily member. Nature *387*, 83-90.

McPherron, A. C., and Lee, S. J. (1997). Double muscling in cattle due to mutations in the myostatin gene. Proc Natl Acad Sci U S A *94*, 12457-12461.

Meng, X., Brodsky, M. H., and Wolfe, S. A. (2005). A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat Biotechnol *23*, 988-994.

Miller, J. C., and Pabo, C. O. (2001). Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition. J Mol Biol *313*, 309-315.

Moens, C. B., Auerbach, A. B., Conlon, R. A., Joyner, A. L., and Rossant, J. (1992). A targeted mutation reveals a role for N-myc in branching morphogenesis in the embryonic mouse lung. Genes Dev *6*, 691-704.

Mohrs, M., Blankespoor, C. M., Wang, Z. E., Loots, G. G., Afzal, V., Hadeiba, H., Shinkai, K., Rubin, E. M., and Locksley, R. M. (2001). Deletion of a coordinate regulator of type 2 cytokine expression in mice. Nat Immunol *2*, 842-847.

Morris, J. R., Petrov, D. A., Lee, A. M., and Wu, C. T. (2004). Enhancer choice in cis and in trans in Drosophila melanogaster: role of the promoter. Genetics *167*, 1739-1747.

Morris, K. H., Ishikawa, A., and Keightley, P. D. (1999). Quantitative trait loci for growth traits in C57BL/6J x DBA/2J mice. Mamm Genome *10*, 225-228.

Munkhtulga, L., Nakayama, K., Utsumi, N., Yanagisawa, Y., Gotoh, T., Omi, T., Kumada, M., Erdenebulgan, B., Zolzaya, K., Lkhagvasuren, T., and Iwamoto, S. (2007). Identification of a regulatory SNP in the retinol binding protein 4 gene associated with type 2 diabetes in Mongolia. Hum Genet *120*, 879-888.

Murashima-Suginami, A., Takahashi, K., Sakata, T., Tsukamoto, H., Sugai, M., Yanagita, M., Shimizu, A., Sakurai, T., Slavkin, H. C., and Bessho, K. (2008). Enhanced BMP signaling results in supernumerary tooth formation in USAG-1 deficient mouse. Biochem Biophys Res Commun *369*, 1012-1016.

Murphy, M., Stinnakre, M. G., Senamaud-Beaufort, C., Winston, N. J., Sweeney, C., Kubelka, M., Carrington, M., Brechot, C., and Sobczak-Thepot, J. (1997). Delayed early embryonic lethality following disruption of the murine cyclin A2 gene. Nat Genet *15*, 83-86.

Nair, S. K., and Burley, S. K. (2003). X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors. Cell *112*, 193-205.

Nakayama, K., Ishida, N., Shirane, M., Inomata, A., Inoue, T., Shishido, N., Horii, I., Loh, D. Y., and Nakayama, K. (1996). Mice lacking p27(Kip1) display increased body size, multiple organ hyperplasia, retinal dysplasia, and pituitary tumors. Cell *85*, 707-720.

Nijhout, H. F., and Emlen, D. J. (1998). Competition among body parts in the development and evolution of insect morphology. Proc Natl Acad Sci U S A *95*, 3685-3689.

Nirenberg, M. W. (1963). The genetic code. II. Sci Am *208*, 80-94.

Noyes, M. B., Meng, X., Wakabayashi, A., Sinha, S., Brodsky, M. H., and Wolfe, S. A. (2008). A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. Nucleic Acids Res *36*, 2547-2560.

Nurse, P. (1991). Cell cycle. Checkpoints and spindles. Nature *354*, 356-358.

Oikawa, T., and Yamada, T. (2003). Molecular biology of the Ets family of transcription factors. Gene *303*, 11-34.

Oliver, E. R., Saunders, T. L., Tarle, S. A., and Glaser, T. (2004). Ribosomal protein L24 defect in belly spot and tail (Bst), a mouse Minute. Development *131*, 3907-3920.

Oliver, T. G., Grasfeder, L. L., Carroll, A. L., Kaiser, C., Gillingham, C. L., Lin, S. M., Wickramasinghe, R., Scott, M. P., and Wechsler-Reya, R. J. (2003). Transcriptional profiling of the Sonic hedgehog response: a critical role for N-myc in proliferation of neuronal precursors. Proc Natl Acad Sci U S A *100*, 7331-7336.

Ortega, S., Prieto, I., Odajima, J., Martin, A., Dubus, P., Sotillo, R., Barbero, J. L., Malumbres, M., and Barbacid, M. (2003). Cyclin-dependent kinase 2 is essential for meiosis but not for mitotic cell division in mice. Nat Genet *35*, 25-31.

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

Palin, K. (2007) Computational methods for locating and analyzing conserved gene regulatory DNA elements, Ph.D. thesis, University of Helsinki, Helsinki.

Palin, K., Taipale, J., and Ukkonen, E. (2006). Locating potential enhancer elements by comparative genomics using the EEL software. Nat Protoc *1*, 368-374.

Panne, D., Maniatis, T., and Harrison, S. C. (2007). An atomic model of the interferon-beta enhanceosome. Cell *129*, 1111-1123.

Park, H. B., Jacobsson, L., Wahlberg, P., Siegel, P. B., and Andersson, L. (2006). QTL analysis of body composition and metabolic traits in an intercross between chicken lines divergently selected for growth. Physiol Genomics *25*, 216-223.

Pavletich, N. P., and Pabo, C. O. (1993). Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. Science *261*, 1701-1707.

Pelengaris, S., Khan, M., and Evan, G. (2002). c-MYC: more than just a matter of life and death. Nat Rev Cancer *2*, 764-776.

Perillo, B., Ombra, M. N., Bertoni, A., Cuozzo, C., Sacchetti, S., Sasso, A., Chiariotti, L., Malorni, A., Abbondanza, C., and Avvedimento, E. V. (2008). DNA oxidation as triggered by H3K9me2 demethylation drives estrogen-induced gene expression. Science *319*, 202-206.

Perkins, A. C., Gaensler, K. M., and Orkin, S. H. (1996). Silencing of human fetal globin expression is impaired in the absence of the adult beta-globin gene activator protein EKLF. Proc Natl Acad Sci U S A *93*, 12267-12271.

Polite, B. N., Dignam, J. J., and Olopade, O. I. (2006). Colorectal cancer model of health disparities: understanding mortality differences in minority populations. J Clin Oncol *24*, 2179-2187.

Potter, C. J., and Xu, T. (2001). Mechanisms of size control. Curr Opin Genet Dev *11*, 279-286.

Poy, F., Lepourcelet, M., Shivdasani, R. A., and Eck, M. J. (2001). Structure of a human Tcf4-beta-catenin complex. Nat Struct Biol *8*, 1053-1057.

Raff, M. C. (1992). Social controls on cell survival and cell death. Nature *356*, 397-400.

Rahimov, F., Marazita, M. L., Visel, A., Cooper, M. E., Hitchler, M. J., Rubini, M., Domann, F. E., Govil, M., Christensen, K., Bille, C*., et al.* (2008). Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. Nat Genet *40*, 1341-1347.

Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E. D. (2002). Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. BMC Bioinformatics *3*, 30.

Rane, S. G., Cosenza, S. C., Mettus, R. V., and Reddy, E. P. (2002). Germ line transmission of the Cdk4(R24C) mutation facilitates tumorigenesis and escape from cellular senescence. Mol Cell Biol *22*, 644-656.

Remenyi, A., Scholer, H. R., and Wilmanns, M. (2004). Combinatorial control of gene expression. Nat Struct Mol Biol *11*, 812-815.

Renehan, A. G., Booth, C., and Potten, C. S. (2001). What is apoptosis, and why is it important? Bmj *322*, 1536-1538.

Richmond, T. J. (2006). Genomics: predictable packaging. Nature *442*, 750-752.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., *et al.* (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods *4*, 651-657.

Rock, R., Schrauth, S., and Gessler, M. (2005). Expression of mouse dchs1, fjx1, and fat-j suggests conservation of the planar cell polarity pathway identified in Drosophila. Dev Dyn *234*, 747-755.

Rother, K. I., and Accili, D. (2000). Role of insulin receptors and IGF receptors in growth and development. Pediatr Nephrol *14*, 558-561.

Roulet, E., Busso, S., Camargo, A. A., Simpson, A. J., Mermod, N., and Bucher, P. (2002). High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. Nat Biotechnol *20*, 831-835.

Saburi, S., Hester, I., Fischer, E., Pontoglio, M., Eremina, V., Gessler, M., Quaggin, S. E., Harrison, R., Mount, R., and McNeill, H. (2008). Loss of Fat4 disrupts PCP signaling and oriented cell division and leads to cystic kidney disease. Nat Genet *40*, 1010-1015.

Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. Development *132*, 797-803.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W., and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res *32*, D91-94.

Sansom, O. J., Meniel, V. S., Muncan, V., Phesse, T. J., Wilkins, J. A., Reed, K. R., Vass, J. K., Athineos, D., Clevers, H., and Clarke, A. R. (2007). Myc deletion rescues Apc deficiency in the small intestine. Nature *446*, 676-679.

Santamaria, D., Barriere, C., Cerqueira, A., Hunt, S., Tardy, C., Newton, K., Caceres, J. F., Dubus, P., Malumbres, M., and Barbacid, M. (2007). Cdk1 is sufficient to drive the mammalian cell cycle. Nature *448*, 811-815.

Satyanarayana, A., Berthet, C., Lopez-Molina, J., Coppola, V., Tessarollo, L., and Kaldis, P. (2008). Genetic substitution of Cdk1 by Cdk2 leads to embryonic lethality and loss of meiotic function of Cdk2. Development *135*, 3389-3400.

Shioi, T., Kang, P. M., Douglas, P. S., Hampe, J., Yballe, C. M., Lawitts, J., Cantley, L. C., and Izumo, S. (2000). The conserved phosphoinositide 3-kinase pathway determines heart size in mice. Embo J *19*, 2537-2548.

Simpson, P., and Ayyar, S. (2008). Evolution of cis-regulatory sequences in Drosophila. Adv Genet *61*, 67-106.

Smale, S. T., and Kadonaga, J. T. (2003). The RNA polymerase II core promoter. Annu Rev Biochem *72*, 449-479.

Sordella, R., Classon, M., Hu, K. Q., Matheson, S. F., Brouns, M. R., Fine, B., Zhang, L., Takami, H., Yamada, Y., and Settleman, J. (2002). Modulation of CREB activity by the Rho GTPase regulates cell and organism size during mouse embryonic development. Dev Cell *2*, 553-565.

Spilianakis, C. G., Lalioti, M. D., Town, T., Lee, G. R., and Flavell, R. A. (2005). Interchromosomal associations between alternatively expressed loci. Nature *435*, 637-645.

Spitz, F., and Duboule, D. (2008). Global control regions and regulatory landscapes in vertebrate development and evolution. Adv Genet *61*, 175-205.

Spitz, F., Gonzalez, F., and Duboule, D. (2003). A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. Cell *113*, 405-417.

Stanger, B. Z. (2008). The biology of organ size determination. Diabetes Obes Metab *10 Suppl 4*, 16-22.

Stanger, B. Z., Tanaka, A. J., and Melton, D. A. (2007). Organ size is limited by the number of embryonic progenitor cells in the pancreas but not the liver. Nature *445*, 886-891.

Stanton, B. R., Perkins, A. S., Tessarollo, L., Sassoon, D. A., and Parada, L. F. (1992). Loss of N-myc function results in embryonic lethality and failure of the epithelial component of the embryo to develop. Genes Dev *6*, 2235-2247.

Steidl, U., Steidl, C., Ebralidze, A., Chapuy, B., Han, H. J., Will, B., Rosenbauer, F., Becker, A., Wagner, K., Koschmieder, S.*, et al.* (2007). A distal single nucleotide polymorphism alters long-range regulation of the PU.1 gene in acute myeloid leukemia. J Clin Invest *117*, 2611-2620.

Stone, K. C., Mercer, R. R., Gehr, P., Stockstill, B., and Crapo, J. D. (1992). Allometric relationships of cell numbers and size in the mammalian lung. Am J Respir Cell Mol Biol *6*, 235-243.

Strom, A., Bonal, C., Ashery-Padan, R., Hashimoto, N., Campos, M. L., Trumpp, A., Noda, T., Kido, Y., Real, F. X., Thorel, F., and Herrera, P. L. (2007). Unique mechanisms of growth regulation and tumor suppression upon Apc inactivation in the pancreas. Development *134*, 2719-2725.

Sutter, N. B., Bustamante, C. D., Chase, K., Gray, M. M., Zhao, K., Zhu, L., Padhukasahasram, B., Karlins, E., Davis, S., Jones, P. G.*, et al.* (2007). A single IGF1 allele is a major determinant of small size in dogs. Science *316*, 112-115.

Suzuki, Y., Yamashita, R., Shirota, M., Sakakibara, Y., Chiba, J., Mizushima-Sugano, J., Nakai, K., and Sugano, S. (2004). Sequence comparison of human and mouse genes

reveals a homologous block structure in the promoter regions. Genome Res *14*, 1711-1718.

Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., Penegar, S., Chandler, I., Gorman, M., Wood, W.*, et al.* (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet *39*, 984-988.

Trumpp, A., Refaeli, Y., Oskarsson, T., Gasser, S., Murphy, M., Martin, G. R., and Bishop, J. M. (2001). c-Myc regulates mammalian body size by controlling cell number but not cell size. Nature *414*, 768-773.

Tsutsui, T., Hesabi, B., Moons, D. S., Pandolfi, P. P., Hansel, K. S., Koff, A., and Kiyokawa, H. (1999). Targeted disruption of CDK4 delays cell cycle entry with enhanced p27(Kip1) activity. Mol Cell Biol *19*, 7011-7019.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. Nat Rev Genet *10*, 252-263.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A.*, et al.* (2001). The sequence of the human genome. Science *291*, 1304-1351.

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F.*, et al.* (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854-858.

Wallace, J. A., and Felsenfeld, G. (2007). We gather together: insulators and genome organization. Curr Opin Genet Dev *17*, 400-407.

Wang, W., Xue, Y., Zhou, S., Kuo, A., Cairns, B. R., and Crabtree, G. R. (1996). Diversity and specialization of mammalian SWI/SNF complexes. Genes Dev *10*, 2117-2130.

Warren, C. L., Kratochvil, N. C., Hauschild, K. E., Foister, S., Brezinski, M. L., Dervan, P. B., Phillips, G. N., Jr., and Ansari, A. Z. (2006). Defining the sequence-recognition profile of DNA-binding molecules. Proc Natl Acad Sci U S A *103*, 867-872.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P.*, et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Wei, C. L., Wu, Q., Vega, V. B., Chiu, K. P., Ng, P., Zhang, T., Shahab, A., Yong, H. C., Fu, Y., Weng, Z.*, et al.* (2006). A global map of p53 transcription-factor binding sites in the human genome. Cell *124*, 207-219.

West, A. G., and Fraser, P. (2005). Remote control of gene transcription. Hum Mol Genet *14 Spec No 1*, R101-111.

Wolfe, S. A., Grant, R. A., Elrod-Erickson, M., and Pabo, C. O. (2001). Beyond the recognition code": structures of two Cys2His2 zinc finger/TATA box complexes. Structure *9*, 717-723.

Woods, K. A., Camacho-Hubner, C., Savage, M. O., and Clark, A. J. (1996). Intrauterine growth retardation and postnatal growth failure associated with deletion of the insulin-like growth factor I gene. N Engl J Med *335*, 1363-1367.

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K.*, et al.* (2005). Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol *3*, e7.

Wu, L., de Bruin, A., Saavedra, H. I., Starovic, M., Trimboli, A., Yang, Y., Opavska, J., Wilson, P., Thompson, J. C., Ostrowski, M. C.*, et al.* (2003). Extra-embryonic function of Rb is essential for embryonic development and viability. Nature *421*, 942-947.

Wu, L., Timmers, C., Maiti, B., Saavedra, H. I., Sang, L., Chong, G. T., Nuckolls, F., Giangrande, P., Wright, F. A., Field, S. J.*, et al.* (2001). The E2F1-3 transcription factors are essential for cellular proliferation. Nature *414*, 457-462.

Yang, Y. (2009). Growth and patterning in the limb: signaling gradients make the decision. Sci Signal *2*, pe3.

Yoshida, H., Kong, Y. Y., Yoshida, R., Elia, A. J., Hakem, A., Hakem, R., Penninger, J. M., and Mak, T. W. (1998). Apaf1 is required for mitochondrial pathways of apoptosis and brain development. Cell *94*, 739-750.

Zanke, B. W., Greenwood, C. M., Rangrej, J., Kustra, R., Tenesa, A., Farrington, S. M., Prendergast, J., Olschwang, S., Chiang, T., Crowdy, E.*, et al.* (2007). Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet *39*, 989-994.

Zhu, C., Byers, K. J., McCord, R. P., Shi, Z., Berger, M. F., Newburger, D. E., Saulrieta, K., Smith, Z., Shah, M. V., Radhakrishnan, M.*, et al.* (2009). High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res.

Zindy, F., Knoepfler, P. S., Xie, S., Sherr, C. J., Eisenman, R. N., and Roussel, M. F. (2006). N-Myc and the cyclin-dependent kinase inhibitors p18Ink4c and p27Kip1 coordinately regulate cerebellar development. Proc Natl Acad Sci U S A *103*, 11579-11583.