BIOINFORMATIC ANALYSIS OF EPITHELIAL:MESENCHYMAL CROSSTALK DURING MOUSE GUT DEVELOPMENT AND PATTERNING

by

Xing Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2009

Doctoral Committee:

Professor Deborah L. Gumucio, Chair
Professor Andrzej A. Dlugosz
Professor David J. States
Associate Professor Kerby A. Shedden
Assistant Professor Zhaohui S. Qin

2009

DEDICATION

To my wife and my son

ACKNOWLEDGEMENTS

I would like to extend my sincere thanks to my advisor, Deborah Gumucio, for her full support, excellent training, amazing insights, incredible patience and inspiration, and great passion for science. Her enduring zeal and persistent diligence set a wonderful example for me to follow as a scientist and person. Without her complete support and encouragement, this thesis work was impossible. I have learned a great deal from her. Dr. Gumucio has always been very supportive and considerate in many aspects of my family, especially for my son's birth. I cannot give enough thanks to her. It has been a great pleasure having her as my mentor and I am grateful for this invaluable opportunity.

I am also grateful to other members of my thesis committee: Drs. Andrzej A. Dlugosz, Zhaohui Qin, Kerby Shedden and David States. I really appreciated all of your incredible support, advice and guidance. Dr. States has been my co-advisor and offered excellent insights and direction both in bioinformatics and biology fields； I also sincerely thank Dr. States for his invaluable support and motivation, especially in the first couple years when I arrived at The University of Michigan. Dr. Shedden was the mentor of my first rotation. He gave me excellent advice and direction on microarray data analysis, statistics and computing programming with great patience. Dr. Qin has been a very supportive and talent advisor on microarray data mining, statistical analysis, and many other fields. I have learned a lot from him, especially for microarray data analysis, transcription factor binding site analysis, clustering, and

programming in R and BioConductor. Dr. Dlugosz has been offering me outstanding advice and inspiration Hedgehog patterning and its roles in controlling gut development. His helpful comments in my Gut group presentations were greatly appreciated. It is truly an honor to have such an excellent committee.

I also offer my gratitude to the past and present members in the Gumuco lab. I could not imagine being able to work with more enthusiastic, dedicated, smart and cooperative people. Specifically, in the epithelium and mesenchyme project, I would like to thank Will Zacharias for teaching me how to make cDNA and perform RT-PCR; Åsa Kolterud for in situ hybridization. In the pyloric border microarray project, I thank Chunbo Hu for in situ hybridization, Tracy Qiao for tissue collection and Sox2 staining, Neil Richards for PCR analysis and probe production. A special thank you to Aaron Udager, who is a co-first author on the border microarray paper. Aaron performed tissue collection, PCR and in situ hybridization. He also helped with data assembly and table formatting. He will continue to study this interesting dataset for his thesis work. I am also grateful to Jierong Long for taking care of the mice and genotyping. Thanks to Kate Walton for the help with formatting problems; you really save me. Thanks to Mike Czrewinski and other previous lab managers for helping me with ordering and other issues. I also extend my gratitude to the rest of the Gumucio lab, Ann Grosse, Andrea Waite. Thank you all for your inspiring discussions, kind help, and for creating such a productive, enjoyable and friendly atmosphere. I have so many great memories of fun times that we shared in the lab, potlucks, Christmas parties, North Michigan party, etc. I really could not have asked for a more wonderful group of labmates.

Many thanks are extended to Becky Pinter in Center for the Organogenesis for her

kind help with many things. My thanks also to Yili Chen, Ji Chen and other members in the States Lab for the great discussions and fun interactions. I thank the Gut Group for inspiring discussions. I also thank the UM Cancer Center Microarray Core personnel for performing an microarray hybridizations. My gratitude is also given to the staff in the Bioinformatics Program, CCMB and CDB for their complete support.

In particular, I am extremely grateful for my beautiful wife, Qingling, and my lovely son, Ruihao, for their love, happiness and complete support during my Ph.D study. I also greatly appreciate my parents, parents-in-law, brothers, sisters-in-law and brother-in-law for their love, constant support and encouragement.

Thank you all.

Table of Contents

List of Figures

Figure

List of Tables

Table

ABSTRACT

BIOINFORMATIC ANALYSIS OF EPITHELIAL:MESENCHYMAL CROSSTALK DURING MOUSE GUT DEVELOPMENT AND PATTERNING

By

Xing Li

Chair: Deborah L. Gumucio

The small intestine develops from a tube of endoderm wrapped by mesoderm. Crosstalk between the endodermally derived epithelium and the underlying mesenchyme is required for regional patterning and proper differentiation of the developing intestine. In this thesis, microarray technology was combined with bioinformatics techniques to study two aspects of small intestinal organogenesis. First, the transcriptomes of the separate mesenchymal and epithelial compartments of the perinatal mouse intestine were examined. It was found that the vast majority of soluble inhibitors and modulators of signaling pathways such as Hedgehog, BMP, Wnt, Fgf, and Igf are expressed predominantly or exclusively by the mesenchyme, accounting for its known ability to dominate instructional crosstalk. Additionally, while epithelially enriched genes tended to be highly tissue restricted in their

expression pattern, mesenchymally enriched genes were broadly expressed in multiple tissues. Thus, the unique tissue-specific signature that characterizes the intestinal epithelium is instructed and supported by a mesenchyme that itself expresses genes that are largely non-tissue specific.

In a second study, gene expression profiles were analyzed during the formation of the pyloric border. At E14.5, before this border is established, gene expression patterns in stomach and nearby duodenum were similar. However, at E16.5, border formation was accompanied by the up-regulation of about 2000 genes specifically in the duodenum. Combining the results from these two microarray experiments revealed that >95% of up-regulated genes were epithelial. This work establishes for the first time that epithelial border formation occurs via a massive change in duodenal (not stomach) character. Genes that are specifically expressed at the border (Nkx2.5, Gata3, nephrocan) and might be involved in border specification were identified, as were transcription factors (Hnf4α, Hnf4γ, Tcfec, Creb3l3, etc.) that are likely to be important in establishment of intestinal identity, a process herein called "intestinalization". Taken together, the results of these studies provide new insights into tissue crosstalk and the specific transcriptional networks that are responsible for intestinal organogenesis.

CHAPTER I

INTRODUCTION

The biology of gut development

Brief introduction to gut development

The gastrointestinal (GI) tract plays a critical role in digestion and nutrient absorption. This GI system extends from mouth to anus and includes the GI tract proper as well as associated solid organs (pancreas, liver, etc). The embryonic gut tube in vertebrates is composed of endodermal epithelium and splanchnic mesodermal mesenchyme. In the process of embryonic development, the gut tube is properly divided into different regions that eventually develop into the various highly specialized and morphologically different organs, including esophagus, stomach, duodenum, small intestine, and large intestine. Their morphology and characters are directly associated with their distinct functions, such as food delivery, food storage, digestion, absorption, or waste packing and excretion. Understanding the fundamental processes of alimentary tract organogenesis during embryonic development has the potential to shed light on the mechanisms of disease in these tissues, since often, these diseases involve reactivation of the embryonic program or re-use of embryonic signaling pathways.

At gastrulation, the cells of developing embryo are divided into three principal germ layers: endoderm, mesoderm, and ectoderm. The gut epithelium is derived from

endoderm, which is initially the ventral-most, and later the innermost layer of the vertebrate embryo. This gut endoderm is surrounded by mesoderm, which becomes remodeled during development into muscle, blood vessels, lymphatics and other supporting tissues.

The development of the gastrointestinal tract in vertebrates can be divided into four fundamental stages (Wells & Melton 1999): (i) formation of a sheet of endoderm supported by lateral plate mesoderm immediately after during gastrulation; (ii) morphogenesis of this bi-layered sheet of tissue into the primitive gut tube; (iii) establishment of organ-specific domains (esophagus, stomach, small intestine, large intestine, etc) within this primitive tube; (iv) organ-specific differentiation of each domain. The fundamental mechanisms that regulate this entire organogenesis process are complex and many of them are still unexplored.

The formation of endodermal organs begins at E7.5-E8.5 in the mouse embryo. At the end of gastrulation (E7.5), the endoderm is a one cell layer thick cup of approximately 500 cells, which forms the ventral-most surface of the embryo. Within the following 24 hours (E8.5), a series of morphogenetic movements remodel this one cell layer endoderm cup and finally transform it into a primitive tube. The signals and pathways that establish these cell movements are under intense study. However, it is clear that even at this early time, the forming tube is somewhat patterned along its anterior/posterior and dorsal/ventral axis. This is believed to occur mainly through recruitment of specific pre-patterned mesenchymal populations to the forming endodermal tube at precise locations. The pattern is then played out via bidirectional mesenchymal/epithelial crosstalk.

Despite some regional patterning in both mesoderm and endoderm, the primitive endodermal tube appears initially morphologically homogeneous along its length. Anterior-posterior (A/P) patterning of different organs along the tube is induced by mesodermal signals (Biben et al 1998b, Lawson & Pedersen 1987).

The early anterior and posterior endoderm express different markers, with Hesx1 (Thomas et al 1998) and Otx2 (Ang et al 1996, Perea-Gomez et al 2001, Rhinn et al 1998) in anterior endoderm and IFABP or intestinal fatty acid binding protein (Green et al 1992), and Cdx2 in the posterior region (Beck et al 1995, Fang et al 2006). Epstein et al showed that Otx and Cdx family genes not only mark the anterior and posterior boundaries but also are required for establishing early tissue pattern in the frog (Epstein et al 1997) though their functions in mouse endoderm A/P specification is not clear.

As the pattern plays out, foregut endoderm contributes to the formation of the esophagus, lung, stomach, liver, and pancreas. The midgut contributes to formation of the small intestine, and the hindgut contributes to formation of the cecum and large intestine (Wells & Melton 1999, Zorn & Wells 2007).

At the same time of tube patterning, buds of endoderm and associated mesoderm also form from this primitive tube at E9.5 to E10.5 and undergo organ specific differentiation. These buds eventually will develop into gut-associated organs, such as liver, pancreas, thymus, etc. The signals determining and patterning these endodermal buds derive from adjacent structures of both mesodermal and ectodermal origin and seem to be both inductive and permissive. For example, the heart (cardiac mesoderm) provides an instructive signal to the hepatic bud while endothelial cells of the dorsal

aorta provide permissive signals for dorsal pancreas development (Wells & Melton 1999).

The establishment of vertebrate gut tube A-P and D-V pattern requires expression of many transcription factors, of which the homeodomain containing factors may be especially important (Beck 2002, Beck 2004, Beck et al 2000, Bort et al 2006, Choi et al 2006, Kim et al 2007b, Playford 2002, Zacchetti et al 2007). The anterior gut tube expresses several Hoxb genes (Huang et al 1998), Nkx2.6 (Biben et al 1998a, Nikolova et al 1997, Tanaka et al 2000) Nkx2.1 (also called TTF1/NKX2A/BCH) (Maeda et al 2006, Minoo et al 1995, Reynolds et al 2003, Rossi et al 1995), Pax8 (Mansouri et al 1998) and Pax9 (Peters et al 1998). The region of the gut tube that contributes to stomach, pancreas, and duodenum expresses homeodomain factors, such as Isx (Choi et al 2006), Nkx2.2 (Desai et al 2008, Doyle & Sussel 2007, Sussel et al 1998), Isl-1 (Ahlgren et al 1997), Pdx1 (Ahlgren et al 1996, Boucher et al 2008, Svensson et al 2007), Pax4 (Larsson et al 1998, Ritz-Laser et al 2002, Sosa-Pineda et al 1997) and Pax6 (Liu et al 2003, Martin et al 2004, St-Onge et al 1997, Trinh et al 2003), while the posterior gut tube expresses many genes in the Hoxd cluster (Roberts et al 1995), as well as Cdx1 and Cdx2 (Beck 2002, Beck 2004, Beck et al 1995, Beck et al 2000).

Epithelial-mesenchymal crosstalk

The gut tube is a two-layered organ. Endoderm is surrounded by mesoderm. Regional gut tube patterning and gut-derived organogenesis along the anterior-posterior axis require bi-directional endoderm-mesoderm crosstalk. The epithelium is in a constant crosstalk with the underlying mesenchyme to maintain or regulate stem cell activity, proliferation in transit-amplifying compartments, lineage commitment, differentiation,

and cell death. These reciprocal interactions are carried out by soluble signals that pass between of the epithelium and the underlying mesenchyme (Roberts 2000, Wells & Melton 1999).

Recombination experiments in which epithelium and mesenchyme from different gut regions are separated and recombined in grafts (e.g. intestinal mesenchyme and stomach epithelium) show that mesenchyme plays a major instructional role in epithelial differentiation. Thus, for example, the intestinal mesenchyme can instruct stomach epithelium to become intestinal epithelium (Kedinger et al 1998b). The intestinal epithelium is also highly instructive, but its instructive power is limited to a short developmental window of time. The exact signals that are responsible for these instructions are still being investigation, but the work shown in chapter 2 of this thesis begins to shed some light on this question.

Regional patterning of the gut tube establishes the organ domains. Then, each organ (eg, stomach, intestine) begins to differentiate in an organ-specific way. In the intestine, this differentiation process involves remodeling of the simple gut tube into villi and crypts. Before villus formation, the epithelium is of the stratified squamous type. At E14.5, this multilayered squamous epithelium undergoes remodeling to become a single layer of columnar epithelium that will finally form villi at E16.5 (Calvert & Pothier 1990, Mathan et al 1976). The villi are finger-like structures lined by absorptive epithelium that projects into the lumen; epithelial cells of the villi execute the function of digestion and nutrient absorption after birth. The inter-villus regions of small intestine, an area where proliferative cells concentrated, will be remodeled to form flask-shaped crypts. The lower part of the intestinal crypt provides the niche for the intestinal stem cells. These stem cells constantly give rise to new

epithelial cells and this proliferative process pushes the epithelial cells up and out of the crypt and onto the villi. Cells stop proliferating at the crypt-villus junction and begin to differentiate. They continue to migrate up to the top of the villi, where they are finally sloughed off into the lumen. In this way, the entire epithelial surface is renewed every 4 days.

There are four types of intestinal epithelial cells: enteroendocrine cells, enterocytes, goblet cells and Paneth cells (Sancho et al 2004). While the Paneth cells reside at the base of the crypts near the stem cells and secrete antimicrobial peptides and enzymes such as cryptidins, defensins, and lysozyme (Porter et al 2002), the other three cell types undergo the constant crypt-to-tip migration discussed above. Enterocytes secrete hydrolases and absorb nutrients; enteroendocrine cells which are rare secrete hormones including serotonie, substance P, and secretin; and goblet cells produce a protective mucous lining (Hocker & Wiedenmann 1998). Though cell lineage decisions are thought to involve Notch signaling among epithelial cells, the possible role of mesenchyme in this process is poorly studied.

Specific pathways important in intestinal development

A complete understanding of the process of intestinal development is not currently available, but several studies have begun to reveal specific roles for several signaling pathways, including Hedgehog, Wnt, Bmp, Notch, TGF-beta and Fgf (Sancho et al 2004).

Most of these signaling molecules (except Notch) are soluble. In order to gain a better understanding of GI tract development, it is critical to know not only which cells or tissues (i.e., epithelium or mesenchyme) express these factors or signals but also

which cells receive those signals. Typically, this has been established by *in-situ* hybridization or by immunohistochemistry. However, as it will be demonstrated in this thesis, microarray approaches can give a valuable global picture. Below, I will discuss some of the major functions that have been determined for some of the most important signaling pathways in order to provide perspective for the interpretation of my own studies.

Hedgehog pathway

The Hedgehog signaling pathway plays vital roles in vertebrate gastrointestinal development in the embryo and in homeostasis in adult life; perturbation in Hedgehog signaling results in diseases or malformations of the gastrointestinal tract and other tissues. (Farzan et al 2008, Fukuda & Yasugi 2002, Lees et al 2005, Omenetti & Diehl 2008, Parkin & Ingham 2008, van den Brink 2007, Zavros 2008)

There are three Hedgehog (Hh) proteins in vertebrates: Indian Hedgehog (Ihh), Sonic Hedgehog (Shh), and Desert Hedgehog (Dhh). Both Ihh and Shh have important functions in gut development as measured by the severe GI phenotypes seen when either of these two proteins is deleted (Ramalho-Santos et al 2000). However, the Dhh protein is not highly expressed in the intestine and the Dhh knockout has no gut phenotype (Bitgood et al 1996). Patched1 (Ptch1) and Patched2 (Ptch2) are the receptors for Hh proteins. Smoothened (Smo) is a transmembrane protein that does not bind Hh, but is responsible for turning on or off the Hh signal transduction by virtue of its reversible inhibition by Ptch. Three zinc-finger family members, Gli1, Gli2 and Gli3, are downstream transcription factors of the Hedgehog signaling pathway. In the absence of Hh signals, Ptch blocks the activity of Smo. In the absence

of Smo activity, Gli factors are degraded, either partially, to form a repressor, or completely. When the Ptch receptor binds Hh, Smo inhibition is released and this triggers a signaling cascade that results in the translocation of full length Gli factors into the nucleus. After translocation, the Gli factors drive the expression of target genes of Hedgehog signals (Ingham & McMahon 2001, Nybakken & Perrimon 2002). Another important member of Hedgehog pathway is Hedgehog-interacting protein (Hhip). Hhip is a transmembrane protein and can bind Hh proteins with an affinity comparable to the Hedgehog receptor Ptch and hence is an important regulator of Hedgehog signaling pathways. It is important to realize that Hhip, as well as Gli1 and Ptch1 are direct downstream transcriptional targets of the Hh signaling pathway.

Hedgehog signals act as morphogens during development as they control cell fate specification in a concentration-dependent manner. Hedgehog signals are strictly paracrine in intestine as the the receptors (Ptch and Smo), transcription factors (Gli1, Gli2 and Gli3) and regulator (Hhip) are all expressed in intestinal mesenchyme only (Madison et al 2005). Hence, Hedgehog signals (Shh and Ihh) are generated in epithelium and secreted into mesenchyme to have functions there (Madison et al 2005). The Hedgehog signaling pathway is required for the correct formation of intestinal villi (Madison et al 2005), concentric patterning of intestinal smooth muscle (Sukegawa et al 2000), and patterning of enteric neurons (Ramalho-Santos et al 2000, Zhang et al 2001b). Hh over-expression has also been implicated in stomach and colon cancer and this was thought to be due to activation of autocrine Hh signaling in epithelial cells (Nielsen et al 2004, van den Brink et al 2004, Varnat et al 2006). However, a recent study of the role of Hh signaling in colon cancer revealed that in such cancers, Hh signaling is indeed paracrine and epithelial proliferation is driven by

abnormal mesenchymal signals that were induced by too much Hh (Yauch et al 2008).

Wnt pathway

The Wnt signaling pathway is crucial for controlling cellular decisions between proliferation and differentiation in the epithelium (Sancho et al 2004). Beta-catenin is the central member in the Wnt signaling cascade. The stability of Beta-catenin is regulated by the Apc tumor suppressor complex (Kohler et al 2008). Complexes of Frizzled seven-transmembrane molecules and Lrp5/6 are the receptors for Wnt signals. When Wnt signals are absent, caseine kinase 1 and GSK3-beta in the Apc complex phosphorylate several highly conserved Ser/Thr amino acid residues in the N terminus of Beta-catenin. This phosphorylation leads to the ubiquitination and degradation of Beta-catenin. In the absence of Wnt signals, the DNA-binding proteins T cell factor/Lymphoid-enhancing factors (TCF/LEF) occupy Wnt target genes along with co-repressors and repress their expression. When Wnt signals are present, the kinase activity of the Apc complex is blocked and this results in Beta-catenin accumulation. Beta-catenin translocates into the nucleus and engages TCF/LEF and transiently converts TCF factors into transcriptional activators to induce the transcription of TCF target genes (Sancho et al 2004).

Recent studies on epithelial cell lines and mouse models have shown that the role of Wnt signaling is to mediate proliferation of intestinal stem cell populations and differentiation of the Paneth cell population (Andreu et al 2008, Sellin et al 2008). Both Wnt and Notch signaling pathways are important in the differentiation of secretory populations (Clevers 2006, de Lau et al 2007, Gregorieff & Clevers 2005, Pinto & Clevers 2005, Reya & Clevers 2005, van Es et al 2005).

The Wnt signaling pathway has been implicated in the proliferation of intestinal epithelial progenitor cells (Bienz & Clevers 2000, Booth et al 2002, Clarke 2006, Kinzler & Vogelstein 1996). It has been shown that proliferative cells at the bottom of crypts of the small intestine and the colon accumulate nuclear Beta-catenin. As Wnt signals decline, these cells differentiate, and switch from a crypt-like phenotype to a differentiated villus epithelial phenotype (Batlle et al 2002, van de Wetering et al 2002). Loss of TCF4 hastens this process, causing loss of proliferative compartments in the small intestine (Korinek et al 1998). Moreover, overexpression of the soluble Wnt inhibitor Dkk1 dramatically reduces epithelial proliferation and results in the loss of crypts (Gregorieff et al 2005, Pinto et al 2003). Wnt signaling is also likely to play a role in the biology of mesenchymal cells, but this has not been well studied yet.

Bmp pathway

Bone morphogenetic proteins (BMPs) are important in intestinal development especially for epithelial renewal (Ishizuya-Oka & Hasebe 2008, Ishizuya-Oka et al 2006). BMPs bind type I and type II serine-threonine kinase receptors which specifically phosphorylate and activate Smad1, Smad5, and Smad8. Smad1, Smad5 and Smad8 (receptor-regulated samds, termed R-SMADS) associate with Smad4 (common SMAD, co-SMAD) and translocate to the nucleus where this SMAD complex interacts with other transcription co-activators or co-repressors to regulate transcription of downstream target genes. It has been shown that Bmp4 is expressed in the intervillus mesenchyme in the adult tissues (Haramis et al 2004, Hardwick et al 2004). The phosphorylation of the downstream transcrtiption factors, Smad1, Smad5, and Smad8 (Zwijsen et al 2003) in villus epithelium as well as mesenchymal cells is an indication that BMP from the mesenchyme acts as both a paracrine and autocrine

signal. The mutation of BMP receptor type 1A or Smad4 is associated with Juvenile Polyposis Syndrome, a disease that is characterized by the formation of thousands of polyps in the intestine (Howe et al 2001, Howe et al 1998, Sayed et al 2002, Zhou et al 2001). A recent study by He et al concludes that this syndrome is caused by loss of Bmpr1a in the epithelium. However, that study used a Cre driver to delete Bmpr1a that is active in both epithelium and mesenchyme. Thus it was still not clear whether the critical BMP function causing over-proliferation of epithelium is targeted to mesodermal or epidermal cells. Indeed, a specific knockout of Bmpr1a in only the epithelium using a Villin promoter was later shown to affect cell fate choice but not proliferation of intestinal epithelium (Auclair et al 2007). Therefore, it seems clear that a relay is active: BMP signals that are received by mesenchymal cells cause these cells to send a signal to epithelium that says "don't proliferate." It is not clear what this response signal is currently.

Notch pathway

The Notch signaling pathway is essential in cell fate specification and differentiation in the intestinal epithelium, as well as spatial patterning (Artavanis-Tsakonas et al 1999, Sancho et al 2004). All Notch receptors (Notch 1, 2, 3, 4) and five ligands (Delta-like 1, 3, and 4; Jagged 1 and 2) are transmembrane proteins. In the absence of the ligands, the transcriptional factor in the Notch pathway, CBF1/RBPjk, acts as a repressor. When Notch binds to its ligands on an adjacent cells, Notch will be cleaved and the Notch intracellular domain (NICD) will translocate to nucleus to bind its transcription factor CBF1/RBPjk to activate the downstream target genes, one of which is hairy/enhancer of split (HES), a member of the basic helix-loop-helix transcriptional factor family. HES can therefore regulate other downstream target

genes (Baron 2003, Hansson et al 2004, Iso et al 2003, Mumm & Kopan 2000).

Members of the Notch signaling pathway are expressed both in embryonic and adult mouse intestine and are involved in controlling the cell's decision to proliferate or differentiate as well as cell lineages (Jensen et al 2000, Schroder & Gossler 2002). In addition, the cell fate, that is, the choice to be secretary (goblet, enteroendocrine or Paneth) versus absorptive (enterocyte), is controlled by MATH 1, which is a downstream target of HES in intestine. Animals with reduced HES in small intestine have fewer absorptive cells but increased mucous-secreting goblet cells and enteroendocrine cells (Jensen et al 2000, Yang et al 2001).

Pyloric border formation in gut development

One of the interesting morphological aspects of the gut tube is the fact that there are several points along the length of the gut where very distinctive borders are found. Despite the fact that the gut tube is initially (in the fetus) morphologically identical from mouth to anus, the adult tube has clear organ-specific character. Our laboratory has been interested in one of these boundaries between two organs: the pyloric border. This border is of interest for two reasons. First, the mechanisms that control boundaries between tissues are not well understood, but are are of intense interest from a biological and developmental standpoint. Second, this particular boundary is where cells with intestinal identity meet directly with cells of stomach identity. How these cells "know" how to be intestine or stomach is of interest because there are pathological lesions found in the stomach within which cells take on intestinal identity. These lesions with cells of the "wrong" (intestinal) address are called intestinal metaplasias. They are precursers to gastric cancer (Schmidt et al 1999, Silberg et al 2002). Therefore, the understanding of how the stomach and intestine gain and

maintain their identity is of great interest.

In the adult gastrointestinal tract, the morphologic border between stomach and small intestine is literally one cell thick. The patterning mechanisms that underlie the development of this sharp regional division from a once continuous endodermal tube are still obscure. Interestingly, this regional division occurs quite late in fetal life, at E16.5 (Braunstein et al 2002). It is clear that some pattern is laid down early. For example, the region-specific expression of certain genes (e.g., intestine-predominant expression of the actin binding protein villin) can be detected as early as 9.0 days post coitum. However, there is no sharp boundary of villin expression between intestine and stomach until 4-5 days later (Braunstein et al 2002). The refinement in cellular identity needed to finally establish intestine from stomach is a late event which we call "intestinalization".

The previous work in our lab has shown that villin responds to cues for intestinal identity (Braunstein et al 2002, Madison et al 2002). Therefore, we used this gene to determine exactly when the border forms. We found that it forms suddenly at E16 in the mouse. The same border is maintained through to adulthood. The villin promoter as well as a villin$^{lacZ/+}$ knock-in mouse have been valuable tools with which to further investigate the mechanisms underlying the formation of this epithelial border. In this thesis, we have used the information gained from the study of the formation of the pyloric border to set up a microarray analysis. The results of that study revealed novel information about gut tube patterning.

Microarray technology and development

Since much of the work done for this thesis involved microarray analysis, I will now

review the application of this technique in biological settings. I will discuss the major platforms available, as well as the leading techniques for further mining of data from array analysis.

Brief introduction to microarray

After the genome sequencing projects are finished in many organisms, one of the next major challenges is how to use the genome sequence information to understand how these genes are regulated on a time- and tissue-specific level. The cadre of genes that are expressed by one cell at a given tissue and time is called its transcriptome. Several high-throughput techniques have been developed for transcriptome analysis, including microarray. The fundamental rationale of microarray technology is based on DNA complementary hybridization according to Watson-Crick rules. Microarray has wide application in biological, medical and clinical fields, including transcriptional profiling (the main application in this thesis), gene copy number variation study (CGH arrays), resequencing, genotyping, single nucleotide polymorphism study (SNP arrays), DNA-protein interaction or transcription regulation study (ChIP-on-chip), gene discovery (Genome Tiling arrays), etc. Microarray technology, for the first time in the history of biomedical science, makes it possible to study the expressions of thousands of genes simultaneously; even the whole genomic expression profile can be analyzed using some platforms. This high-throughput technology combining with other Bioinformatics techniques provides the first opportunity for scientists to understand genomic expression profile, gene regulation and carry out genome-wide network analysis on a holistic level.

For investigating genome expression profiles, gene expression microarray technology is semi-quantitative. It measures the relative amount of mRNA in the sample; this is

proportional to the relative gene expression level. Currently there are two major types of microarrays in abroad use: high density oligonucleotide arrays (Affymetrix GeneChip) and cDNA arrays.

The Affymetrix GeneChip platform is widely used in academia and industry for expression analysis. It is a high-density short oligonucleotide array (Lipshutz et al 1999, Lockhart et al 1996). The manufacture of Affymetrix GeneChips is based on two techniques: photolithography and solid-phase DNA synthesis (Figure 1.1, (Lipshutz et al 1999)). Affymetrix uses light masks to control synthesis of oligonucleotides on the surface of a silicon chip, a technology developed by the microprocessor industry for making silicon chips for computers. The number of genes that could be detected by this platform is limited by the physical size of the array and the achievable lithographic resolution. Current technology allows for several millions of oligo probes to be synthesized on a few square centimeters (1.28 X 1.28 cm). One of the most intriguing aspects of this platform is the presence of multiple probes per genes and mismatch (MM) probes.

As for the principle of Affymetrix GeneChip, each gene is represented by one or more ProbeSets on the Affymetrix microarray. Each ProbeSet is in turn comprised of 11-20 probe pairs and each probe pair has two 25 base long probes, one perfect match (PM) which matches the gene sequence exactly while the mismatch (MM) probe is same as the perfect match except that the middle base in the probe's 13^{th} position is exchanged with its complement (Figure 1.2). The goal of design of multiple probe pairs is to improve specificity because of their short length. The PM probes are designed to hybridize only with transcripts from the target gene (specific hybridization). However, the hybridization of PM probes to other mRNA, non-specific hybridization is

unavoidable. Therefore, the observed intensities need to be corrected. MM probes are introduced to estimate non-specific binding and their location is adjacent to PM (Lipshutz et al 1999). Theoretically or according to the original purpose of the inclusion of MM, MM should not hybridize to the target genes under highly stringent conditions. In practice, this is not necessarily always true. The intensity values of a large number of MM have been shown to be higher than those of PM (Irizarry et al 2003b). Therefore, it should noted that the use of MM control for deriving the final gene expression levels is not universally agreed upon and some studies have shown that ignoring the mismatch data and adopting some statistical models based on perfect match intensity achieve better performance (Chu et al 2002, Irizarry et al 2003b, Li & Wong 2001). This issue will be discussed further below. In addition, there can also typically a high variation in the intensity values of the 16 or more oligonucleotide probes that constitute a probe set which implies that selecting of different set of probes may give a slightly different measure of abundance. In some cases, this can imply mRNA splicing or processing.

The cDNA microarray technology is another platform that has also been extensively used to monitor the relative levels of expression of thousands of genes simultaneously (Schena et al 1995). The basic process of manufacture of this platform is that a robot spotter is used to spot tiny quantities of probe in solution from a microtiter plate to the surface of a coated glass slide or nitrocellulose or nylon membrane. Therefore, cDNA microarrays are also called spotted arrays. Membranes are most suited to applications where radioactivity is used to label the cDNA, while glass slides only support fluorescence-based detection. The probes spotted on cDNA arrays can be cDNA, PCR products, or oligonucleotides. Each probe is complementary to a unique gene. The

probes are fixed on the surface in a number of ways. The traditional method is by non-specific binding to polylysine-coated slides. The samples and controls usually are labeled by differently fluorescence dyes and competitively hybridize to the same cDNA array. This is why cDNA arrays are also called two-channel arrays. In contrast, Affymetrix GeneChips are single channel arrays since one chip is hybridized by one sample, which means the control/reference and treatment need to be hybridized to different chips. From a data analysis point of view, the main difference is that the cDNA arrays usually generate a dataset of ratios while Affymetrix chips give a relative expression levels.

There are advantages and disadvantages of each type of array. The Affymetrix oligonucleotide array can accommodate higher densities of genes, including some hypothetical genes predicted using computing algorithms which are not represented in cDNA libraries. Even the whole set of genes on certain genome can be accommodated by Affymetrix style arrays. The microarray chips that we used for our experiments in the mouse model is 430 2.0. It contains over 45,000 ProbeSets and investigates about 30,000 genes, which is almost all currently known and predicted mouse genes. Usually, the Affymetrix GeneChips exhibit lower variability from chip to chip. And since Affymetrix is a well designed commercial platform, it facilitates microarray data comparison and integration across different research groups throughout the world. Though cDNA arrays are considerably cheaper than Affymetrix chips and offer more flexibility in array design, the spotting of cDNAs is less uniform than the manufactured Affymetrix chips. Since cDNA microarrays typically have an upper limit of 15,000 elements (and often include fewer than 5,000 spots), they are unable to represent the complete set of genes present in higher eukaryotic genomes.

Consequently, this platform is usually developed for detecting a class of interesting genes specific to certain developmental stage or tissue. Since we use the Affymetrix GeneChip platform in this thesis work, I will focus on this platform in the rest of the introduction.

General procedure for microarray experiments

The microarray experiments can be divided into several steps: experimental design, performing the microarray experiments, low-level microarray data analysis, high-level microarray data mining, and follow-up validation and experiments (Figure 1.3).

Designing and performing microarray experiments

The first task in the design of any microarray experiment is to frame a biological question or a hypothesis which will be studied or tested in the microarray experiments. The aims of the microarray experiment, the accessibility of the sample, the availability of funding, and the current research status of the same research area using microarrays (for easily comparing results from different research groups), etc., are all important considerations. When planning the experiments, there are two types of repetitions: technological and biological. Technological repetitions usually involve using the same biological sample repeatedly (e.g., repeated hybridization). Biological repetitions use different samples from different biological individuals. The latter is much more important in microarray experiments.

Once finishing the experimental design, one needs to prepare samples (tissues, cells, etc) under different treatment and control conditions or other time series points. Then extract mRNA, synthesize cDNA and label the cDNA and hybridize it to microarrays. After hybridization, the array is scanned and intensities are extracted for each feature

on the array (The Affymetrix microarray experiment process is shown in Figure 1.4 a).

Background adjustment and normalization

After performing microarray experiments, we need to process the images to obtain the gene expression values. Basically, preprocessing Affymetrix expression arrays involves three main steps: background adjustment, normalization, and summarization of probe level intensities. In any experiment involving multiple Affymetrix microarrays, there is inevitable inherent noise which is introduced in the process of experiments since the experiments are complicated and involve a number of steps, most of which have the potential to introduce noise. The source of the variations may come from differences in mRNA extraction and cDNA amplification, dye-incorporation, efficiency of hybridization, microarray image-scanning process or experimenter bias. Before we can derive any real significant signals from the microarray data, we need to remove or at least to reduce as much as possible the noise. The processes for reducing the noise or variation within/between the arrays are background adjustment and normalization.

Background Adjustment

The background adjustment used in MAS4.0 and MAS5.0 is called regional adjustment. By this method, each array is divided into a grid of n rectangular regions (default n=16). For each region, the lowest 2% of probe intensities are used to compute a background value for that grid. Then each probe intensity is adjusted based on a weighted average of each of the background values. The weights are dependent on the distance between the probe and the centroid of the grid (Affymetrix 1999,

Affymetrix 2001). This method corrects both PM and MM probes.

A different background adjustment accomplished by a convolution strategy is used by the Robust Multichip Average or RMA method (Bolstad et al 2003, Irizarry et al 2003a, Irizarry et al 2003b). Irizarry et al found there are various problems using the MM probes in the preprocessing stage and proposed a procedure that only uses PM. By this method, the PM intensities are adjusted array by array with a global model for the distribution of probe intensities. This method was motivated by the empirical distribution of probe intensities. In brief, the observed PM probes are modeled as the sum of a Gaussian noise component and an exponential signal component (Irizarry et al 2003a, Irizarry et al 2003b).

Since the RMA background adjustment normalization approach ignores the MM intensities, it sacrifices a degree of accuracy for large gains in precision. Moreover, the global background adjustment in RMA ignores the different propensities of probes to undergo non-specific binding; thus this method may underestimate background. Another method, called GCRMA, considers the sequence characteristics of each probe using the probe sequence information released by the manufacturer (Wu & Irizarry 2004, Wu & Irizarry 2005). In this model, an affinity measure is calculated by using the probe sequence information. A background adjustment method motivated by this model has been implemented and together with quantile normalization and the medial polish procedure used by RMA to define a new expression measure.

Normalization

Following the background adjustment, there are two main types of methods for performing normalization on the probe intensity level. The first type is called the

baseline array method; it selects a baseline array, usually the array with the median of the median intensities. There are scaling methods (Affymetrix 1999, Affymetrix 2001) and non-linear methods (Li & Hung Wong 2001, Li & Wong 2001) to accomplish this. The scaling method is the standard Affymetrix normalization method; it is used both in MAS4.0 and MAS5.0. After selection of the baseline array, all the arrays are normalized to this array by multiplying a scale to adjust them so that they have the same mean intensity as the baseline array. The scale is calculated by dividing the trimmed mean intensity of the baseline array by the trimmed mean intensity of the normalized array (Affymetrix 1999, Affymetrix 2001). This method is equivalent to selecting a baseline array and then fitting a linear regression without an intercept term between each array and the chosen array. The non-linear method performs non-linear adjustments between the arrays and tends to out-perform linear adjustments such as the scaling method. Several non-linear relationships have been proposed including cross-validated splines (Schadt et al 2001) and running median lines (Li & Hung Wong 2001, Li & Wong 2001). For a typical implementation, the normalizing relationship is fitted using a rank-invariant set of points, that is, a set of points that has same rank ordering on each array.

A second group of widely-used normalization methods is called complete data algorithms since this approach combines information from all arrays in one experiment to establish the normalization relationship. Two algorithms in this group are the cyclic loess and contrast based method, both of which are based on the M versus A plot. Here M is the difference in log expression values and A is the average of the log expression values (Astrand 2003, Bolstad et al 2003, Dudoit et al 2002). The third method is Quantile normalization; this method makes the distribution of

probe intensities the same for each array in multiple microarray experiment. This is the normalization method used in the Robust Multichip Average (RMA) method (Bolstad et al 2003).

Summarization of Probe Intensities

Affymetrix high density oligonucleotide arrays rely on multiple different probe pairs for each gene and it is necessary to condense these probe pairs into a single intensity for each gene. This process is called summarization. There are several methods to obtain the gene expression values: Average difference (AvgDiff, MAS4.0)(Affymetrix 1999), the MAS5.0 statistical algorithm (Affymetrix 2001, Hubbell et al 2002), the Model-based Expression Index (MBEI, implemented in dChip software) (Li & Hung Wong 2001, Li & Wong 2001), and the Robust Multichip Average (RMA) (Irizarry et al 2003b) and GCRMA (Wu & Irizarry 2004, Wu & Irizarry 2005). Some of these algorithms use mismatch information to make adjustment for non-specific hybridizations and some completely ignore MM.

In the early version of the Affymetrix software MAS 4.0, an Average Difference between PM and MM probe pairs was calculated. Since the mismatch (MM) was originally designed to detect probe-specific non-specific hybridization, we could adjust the PM probe intensities by subtracting the MM intensities from the corresponding PM intensities. Average Difference was calculated as follows:

AvgDiff = (Sum of (PM – MM) of all probes for each gene) / (number of probe pairs)

Note that probes that deviate by more than three standard deviations from the mean are excluded from the calculation. There are negative or very small AvgDiff values. The possible reasons for these values are either that the target is absent or that there is

non-specific hybridization.

For intensities from a typical Affymetrix microarray experiment, as many as 30 percent of MM probes have higher intensities than their corresponding PM probes (Irizarry et al 2003a, Irizarry et al 2003b, Naef et al 2002). Consequently, when raw MM intensities are subtracted from PM intensities using AvgDiff methods in MAS 4.0, many negative expression values are generated. This becomes problematic and makes little sense because it is impossible for a gene expression value to below zero. Furthermore, the negative values preclude the use of logarithms which are widely used for data transformation in statistical analysis.

In the new version of the Affymetrix software (MAS 5.0), this issue is solved by introducing ideal MisMatch (IdealMM) probes. The values of IdealMM are designed to be smaller than the corresponding PM intensities. The strategy of the IdealMM is to use MM when its intensity is less than PM or a quantity smaller than the PM in other case. This is done by computing the specific background for each ProbeSet which is a robust average of the log ratios of PM to MM for each probe pair in the ProbeSet (Affymetrix 2001, Affymetrix 2002). The value is calculated by:

$$\text{Signal} = \text{TukeyBiweight}(\log(PM_n - \text{IdealMM}_n)),$$

where Tukey biweight is a robust estimator of central tendency. However, the introduction of IdealMM may affect the normality assumption often used in downstream statistical analysis (Giles & Kipling 2003).

Li and Wong (Li & Hung Wong 2001, Li & Wong 2001) designed a method using a Model-Based Expression Index (MBEI). This model takes into account that probe pairs respond differently to changes in gene expression and that the variation between

replicates is also probe pair dependent. It computes a scaling factor which is specific to probe pair (PMn-MMn) by fitting a statistical model to a series of experiments. It has been shown that this model works with or without MM and usually has lower noise when MMs are excluded. The software, dChip, was designed for fitting the model (weighted average difference and weight perfect match) as well as for detecting outliers and obtaining estimates on reliability (http://www.dchip.org). The weight average difference is calculated by:

Signal = (sum of ((PMn-MMn)*(scaling factor)))/(number of probe pairs)

A comparison of Li and Wong's method with Affymetrix's Average Difference method showed that MBEI (dChip) is superior in a realistic experimental setting (Lemon et al 2002). However, this model parameter estimation works best with 10 to 20 chips.

Since MMs have been observed to detect some specific signals, Irizarry et al. designed the Robust Multichip Average (RMA) that is a procedure for computing expression values (Bolstad et al 2003, Irizarry et al 2003a, Irizarry et al 2003b). It consists of three discrete steps: a convolution model-based background correction, a non-linear quantile probe-level normalization, and a robust mutichip summarization method. I have discussed the first two steps above. RMA completely ignores the information of MM probes in all steps of the algorithm. Summarization is done using a robust multichip linear model to fit on a probeset by probeset basis. Specifically, the standard RMA summarization approach is to use median polish to computer expression values:

RMA-signal = Medianpolish(log(PMn – (scaling factor))),

where the scaling factor is specific to probe PMn and is obtained by fitting a statistical model to a series of experiments. There are also other algorithms for performing background adjustment, normalization and summarization of probe intensities. See Table 1 for a summary of the methods discussed here.

There are several studies comparing how the different normalization and summarization algorithms affect the high-level analysis (Bolstad et al 2003, Hoffmann et al 2002, Ma & Qin 2007, Parrish & Spencer 2004, Shedden et al 2005). However, there is no universal agreement as to which method is best. Generally speaking, the sequence based methods or MM-based methods work better for high- and median-expressing genes and worse for low-expressing genes. Furthermore, which algorithm is best depends on the microarray data and no method of normalization or probeset summarization shows any consistent advantages. We have used the RMA method to perform background adjustment, normalization and summarization of gene expression values in this thesis work (Affy package from BioConductor (Irizarry et al 2003b)).

Identification of significantly changed genes

After obtaining the gene expression values, one of the important tasks for a microarray study is to identify the statistically significantly changed genes across different conditions, tissues or cell types. For comparative microarray experiments involving two groups of samples, usually the t-test is used for identification of significantly changed genes and may be combined with fold change. The t-test looks at the mean and variance of the sample and control distributions and calculates the probability that the observed difference in mean occurs when the null hypothesis is

true (no difference of gene expression for the two compared conditions).

When using the t-test, it is often assumed that the variances in sample and control are equal, which allows the sample and control to be pooled for variance estimation. If there is evidence showing that these variances are not equal, one may use Welch's t-test which assumes unequal variances of the two populations. The t-test also assumes that the data are approximately normal or the sample size is not too small. The gene expression data obtained by RMA are log-transformed which makes the data close to normal but there is no guarantee for this assumption. Giles and Kipling (Giles & Kipling 2003) have demonstrated that deviations from the normal distribution are small for most microarray data, except when using Affymetrix's MAS 5.0 software. Furthermore, the t-test is robust to moderate deviations from the normal distribution. Usually the sample size is small in microarray data analysis due to the high cost or the availability of the samples. The lower the number of replicates, the more difficult it will be to estimate the variance.

There are several solutions to this problem. The simplest solution is to take fold change into account for experiments with small sample size and not consider genes that have lower fold change, for example, less than two-fold change in expression. This will guard low p-values that arise from underestimation of variance, and is sort of similar to the algorithm used in Significance Analysis of Microarray (SAM) (Tusher et al 2001), an approach which adds a small constant to the gene-specific variance. The relationship between p-value and fold-change can be visualized by the volcano plot (p-value on y-axis vs fold-change on x-axis) (Jin et al 2001), genes can then be selected based on the distribution of the data points on the volcano plot. Other solutions for low sample size are to base variance estimates not only on a single gene

measurement, but to include variance estimates from the whole population (Baldi & Long 2001, Kerr & Churchill 2007, Lonnstedt & Speed 2002).

Another alternative method for assessing significance without assuming normality is the Wilcoxon/Mann-Whitney rank sum test (nonparametric test). This approach does not use the actual expression values from the microarray experiment, but rather uses their rank relative to each other. However, since the Wilcoxon test does not measure variance, the significance of this result is limited by the number of replicates. Therefore, one may find that Wilcoxon test gives a poor significance for a low number of replications.

If there are three or more than three groups, conditions, or time series in the microarray experiments, the t-test may not be the ideal method since the number of comparisons grows fast if you perform all possible comparisons between all the conditions. ANOVA (analysis of variation) using F-distribution would be an efficient statistical method to find the statistically significant changed genes in such cases (Kerr et al 2000).

There are other linear models for selecting differentially expressed genes such as LIMMA (Linear Models for Microarray Data analysis) (Smyth et al 2005). It is designed for differential expression analysis of microarray data. The central idea is to fit a linear model to the expression vector. The expression data can be log-ratios from two-color microarrays, or log-intensities from Affymetrix microarray. Empirical Bayes and other shrinkage methods are used to for moderating the genewise variances between genes which makes the analyses stable even for experiments with small number of arrays.

There are many software packages for computing t-test and ANOVA. Some of the tools we have used in our microarray data analysis include TM4, which is designed by TIGR (The Institute of Genome Research), and functions and packages within R and BioConductor.

Correction for multiple hypothesis tests

Since microarray data analysis usually contains comparisons of thousands of genes, it is important to consider the effect of multiple hypothesis testing. The p-value of 0.05 (which is frequently used when selecting significantly changed genes) means that you have a probability of 5% of making a type I error (false positive call) on one gene. If there are 10,000 genes on the microarray, you expect 500 type I errors (false positive calls).

For the purpose of correcting multiple hypothesis tests, there are several methods: one-step methods and step-down methods. Both involve using smaller p-values for identifying significant genes by "slashing" the p-value for each test (i.e., gene), so that while the critical p-value for the entire data set might still equal 0.05, each gene will be evaluated at a lower p-value.

The simplest single-step method, known as the Bonferroni correction, divides the alpha value (e.g. 0.05) by the total number of multiple tests (usually this is the total number of genes on the microarray). For example, for 10,000 genes on microarray, the new cutoff is 0.05 divided by 10,000, which is 0.000005. This is a very strict cutoff and using this method, most microarray experiments end with no genes significantly changed.

Step-down methods are less conservative than one-step methods. It makes different

adjustments to the p-values of different genes. This method first ranks the p-value of all the genes in increasing order. Then it compares the p-value with the alpha value after dividing by the total number of tests (N, usually the gene numbers on the array), the second p-value with alpha divided by N-1, the third p-value with alpha divided by N-2, and so on. If the p-value of one gene is less than the corrected alpha value, it is considered significantly changed. Both single-step and step-down methods are generally over-stringent due the large number of tests and only a few genes or no genes at all pass this new cutoff for typical microarray experiments, which results in exclusion of many false negatives.

Another method for correcting the multiple tests in microarray and bioinformatics studies (GO term enrichment and pathway analysis) is false discovery rate (FDR); this works by controlling the proportion of genes that are falsely identified. It can be set to values less strict and will yield a moderate number of genes for study (Benjamini & Hochberg 1995, Dudoit et al 2002, Reiner et al 2003). This method first ranks the genes according to p-values (significance) and then starts at the top of the list and accepts all genes with

$$\text{p-value} <= (i/m)*q,$$

where *i* is the number of genes accepted so far, m is the total number of genes tested and q is the desired FDR. The False Discovery Rate can also be assessed by permutation. This is the method implemented in the software SAM (Tusher et al 2001). It permutes the expression values from sample and control, repeats the t-tests for all genes and gets an estimate of the number of false positives that can be expected at the chosen cutoff (alpha). Then it divides this number by the number of genes that

pass the cutoff on the original un-permuted data to get the FDR.

The final point about multiple hypothesis test correction is not to get overly focused on p-values. Ultimately, what matters is biological relevance and significance. The p-values can help you evaluate the strength of the evidence, and should not be used as an absolute yardstick of significance. Statistical significance is not necessarily the same as biological significance. The lower fold change (FC) and higher p values for some genes, which makes them less statistically significant, may be due to the small sample size (the normal situation in microarray studies) or other reasons. If there are not too many genes passing the FDR control, one may use other methods to understand the gene expression profiles (e.g., Gene Ontology (GO) term enrichment or Gene Set Enrichment Analysis (GSEA) using all the gene expression profiles, discussed below).

Microarray data mining and Bioinformatics

For pairwise comparison of microarray studies, the above analysis will generate a list of regulated genes ranked by the magnitude of up- and down-regulation, and/or ranked by the significance of regulation determined in a t-test. For microarray experiments involving mutifactorial conditions, in attempting to make biological sense of microarray data, it is helpful to visualize the huge data matrix by a method the human brain can process easily. This usually entails graphical representation in the format of line graph or color figures which display genes into clusters with similar expression profiles. A variety of clustering methods have been developed to identify groups of co-regulated genes, including hierarchical clustering, principal component analysis, etc.

Clustering Analysis

Clustering analysis has been using frequently in microarray analysis and can lead to readily interpretable figures (Chu et al 1998, Do & Choi 2008, Eisen et al 1998, Qin 2006, Wang et al 2002). The main application of this method is to identify groups of genes sharing similar expression patterns or profiles and identify spatial or temporal expression patterns, which is often regarded as evidence for similarity of functions, allowing putative annotation of the function of unknown genes. This in turn may imply that the genes are involved in a similar biological process. Consequently, in addition to describing how individual genes respond to certain treatments, microarray analysis can describe the level of coordinate regulation of gene expression on the genome-wide scale. Though not definitive, this type of analysis is at least sufficient to generate hypotheses that can be tested by more traditional molecular biological approaches. Similarity of expression profiles can also imply that the cluster of genes share the same mechanisms of co-regulation. Some clustering methods can even identify both positive and negative regulations in one cluster (Qin 2006). Bioinformatic tools can then be applied to identify upstream regulatory sequences in the promoter regions of these genes which may lead to isolation of transcriptional factors that mediate particular expression profiles. Clustering can be performed both on the genes and the samples (or treatments, mutations, drugs, etc), allowing detection of patterns in two dimensions. In cases where the treatments represent a series of tissue types, drugs or mutations, this two-dimensionality can be an extremely powerful method for identifying similarities in genome-wide responses. Clustering on samples or arrays in tumor microarray studies is essential when seeking new subclasses of tumors or new cell types, which is crucial for successful diagnosis and

treatment. Clustering on genes in cancer studies can identify marker genes that characterize the different tumor classes (feature selection). Furthermore, Clustering can be used for quality control: comparing array/gene clustering results to experimental variables such as array batch, mRNA amplification methods, lab, experimenter, etc.

Clustering involves several distinct steps. First, a suitable distance or similarity between genes must be defined based on gene expression vectors. Then, a clustering algorithm must be selected and applied. The results of a clustering procedure can include both the number of clusters and a set of set of cluster labels for all the genes to be clustered. Appropriate choices will depend on the questions being asked and the data. There are many types of clustering algorithms (Do & Choi 2008, Eisen et al 1998, Qin 2006). One of the frequently used methods is the agglomerative (bottom-up) hierarchical clustering method. The basic algorithm in hierarchical agglomerative clustering is to begin with each data point (gene or sample) as a separate cluster and then iteratively merge the two "closest" clusters until all the data points are in a single big cluster. Here, "close" is defined by clustering criterion which consists of two parts in nonparametric clustering: the similarity or distance measure or metric, which specifies how to compute the distance between two data points; and the linkage, which specifies how to combine these distances to obtain the between-cluster distance since after several steps the clusters may contain more than one data point (sample or gene). In model-based clustering, the clustering criterion is based on the likelihood of the data given the model. The distance or similarity/dissimilarity of different genes can be calculated based on different algorithms, such as Euclidean distance, or L^p distances, cosine vector angle, or one of many other possibilities.

Euclidean distance is defined as:

$$\sqrt{\sum_{i=1}^{n}(a_i - b_i)^2},$$

where *ai*, *bi* are expression data vectors:

$$a = (a_1,...,a_n), b = (b_1,...,b_n)$$

Cosine vector angle is calculated as:

$$1 - \frac{\sum_{i=1}^{n} a_i b_i}{\|a\|\|b\|},$$

where

$$\|a\| = \sqrt{\sum_i^n a_i^2}, \|b\| = \sqrt{\sum_i^n b_i^2}$$

L^p distance, 1 minus Pearson correlation coefficient, is:

$$1 - \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sigma(a)\sigma(b)}$$

where

$$\sigma(a) = \sqrt{\frac{\sum_i^n (a_i - \bar{a})^2}{n}}, \sigma(b) = \sqrt{\frac{\sum_i^n (b_i - \bar{b})^2}{n}}$$

and

$$\bar{a} = \frac{a_1 + ... + a_n}{n}, \bar{b} = \frac{b_1 + ... + b_n}{n}$$

After the distance is computed, the two most similar clusters will be merged in one cluster. Since the clusters may have more than one data point, it is impossible to just look at distance/dissimilarity matrix to determine similarities. This is why the linkage

part is necessary. There are average linkage, single linkage, and complete linkage. Single linkage is defined as the smallest pairwise distance between the members of the two clusters. Average linkage is based on the mean distance of all the possible distances between all the members in the two clusters. Maximum linkage uses the maximum distance in all of the possible distance of the members in the two clusters. In other words, single linkage defines two clusters to be similar if they have at least one pair of similar samples, whereas complete linkage considers two clusters to be similar only if all pairs drawn from the two clusters are similar. Average linkage falls somewhere between the single and complete methods. Complete linkage tends to yield compact clusters, while single linkage can be stringy or elongated. There is no universal agreement on which of these three linkages is best in a given dataset.

A dendrogram or tree is commonly used to visualize the nested structure of clusters resulting from hierarchical clustering, which are usually combined with a heatmap (color-coded matrix) showing gene expression matrix. Clusters or nodes forming lower on the dendrogram are closer together, while upper nodes represent merges of clusters that are farther apart. Since each data point begins as a single cluster in bottom-up hierarchical clustering, the leaves (terminal nodes at the bottom of the dendrogram) each represent one data point (gene or sample), while interior nodes represent clusters of more than one data point. The top node of the dendrogram denotes the entire data set as a single root cluster. Partitions of the tree can be obtained by cutting the tree at different levels of height; closer to the leaves (terminal nodes), yields more clusters. While dendrograms are quite appealing because of their apparent ease of interpretation, they can be misleading. First the dendrogram corresponding to a given hierarchical clustering is not unique since for each merge

one needs to specify which subtree should go on the left and which on the right. So there are many different dendrograms. The default in the R function hclust (cluster package) is to order the subtrees so that the tighter cluster is on the left. Since clustering usually requires vast computing, it is reasonable to filter the data before clustering by removing the genes not changed across different conditions to reduce the required computing burden.

From dendrograms and heatmaps, it is easy to identify groups of genes or samples with similar patterns which might therefore share regulation control. These then can be further investigated to understand the patterns, for example, by searching for common transcription factor binding sites (or motif) analysis in the promoters of the clustered genes. In the follow-up analysis, the biological meaning of clusters can be assessed by their predictive power and their capacity for generating hypothesis by aligning cluster data with Gene Ontology (GO) and other genomic annotation, such as linkage to a literature database.

Principal Components Analysis

The data matrix from microarray gene expression studies is usually highly dimensional and it is impossible to discern any patterns or trends by visual inspection of such a complicated huge matrix. Since visual analysis is usually performed in two or three dimensions, it is necessary to reduce the dimensions of the data matrix before any feasible visual analysis. There are many methods allowing reduction of a matrix of any dimensionality to only two or three dimensions, one of which is Principal Components Analysis (PCA). PCA is a powerful multivariate statistical method for reducing the data dimensions and visualizing the data in a simple x-y graph. Some

important trends or fundamental patterns underlying gene expression profiles can be spotted from the PCA graph alone (Dysvik & Jonassen 2001, Raychaudhuri et al 2000, Xia & Xie 2001) and its popular algorithm is Singular Value Decomposition (SVD) (Alter et al 2000, Alter et al 2003, Holter et al 2000, Wall et al 2001). If replicates are available, it is best to perform PCA on data that has already been filtered for significance. The first principal component captures most of variation in the data and may not necessarily coincide with any of the existing axes. Rather, it will have projections of several or all axes on it. The second principal component captures the maximum amount of variation left in the data and is orthogonal to the first one. Therefore, PCA displays as much of variation in the data as possible in just two or three dimensions. What the analysis will tell you depends on whether there is a trend in the data that is discernible in two or three dimensions. PCA could be carried out on both genes and samples to detect outlier genes in a certain group (such as some group identified in volcano plot) or study the homogeneity of the tissue samples used in the microarray experiments.

In this thesis, we apply PCA on tissue samples in our pyloric border microarray experiments to show that the samples we used in our microarray experiment are quite homogenous and to demonstrate that large changes in gene expression profiles occurred between E14.5 duodenum and E16.5 duodenum.

Transcription factor binding site (TFBS) analysis

Gene expression is regulated in part by transcription factor(s). Transcription factors act by binding to specific DNA elements, *cis* elements or transcription factor binding sites (TFBS), in gene promoter regions. Identification of co-regulated genes and their

transcription factor binding sites (TFBS) are critical issues for understanding gene transcription regulations and networks (Qin et al 2003). Traditionally, methods identify these elements by biochemical and/or molecular genetic procedures (for example gel-shift assays). However, many of these regulatory elements remain to be identified. Bioinformatic strategies together with microarray studies are being employed to help with these issues.

Clustering analysis of microarray data generates groups of genes with similar expression patterns (for example, those having a similar transcription response) in different conditions; this may reflect a commonality in function or a sharing of common regulatory mechanisms. When the genes in one cluster have known and similar function, we can infer the function of unknown genes (or hypothetical genes) within the same cluster.

Given a set of co-regulated genes (genes in the same cluster responding to the same signals or stimuli), it is generally assumed that transcription of most of (if not all) these genes is likely to be regulated by a common transcription factor. Therefore, there may be conserved regulatory sequences in their promoter regions. The most popular strategy is to carry out TFBS analysis on a small group of genes that are identified either by clustering, pathway analysis, or other functional annotation.

The first step in TFBS analysis is to obtain the promoter regions for the group of interesting genes. This requires that the promoter region is defined and included in a database for retrieving. Even when the promoter sequences are available, finding common regulatory elements is inherently complicated because of the degeneracy of such elements. The first problem is that regulatory elements usually are short

(between 6 to 10 bases in length, reflecting the length of helical groove that the proteins or transcription factors recognize through specific hydrogen bonding). The second problem is these regulatory sequences usually are variable to some degree in several locations. The expected frequency of any n-base sequence is one in very 4^n nucleotides which means that it can occur millions of times in the genome. The degeneracy of binding sites (some bases are variable which makes actual binding sites diverge from the canonical motif) only exacerbates this problem. Thus, the problem of identifying *cis* elements that are overrepresented in the promoters of just a small group of genes is acute. Furthermore, transcriptional regulation is complicated. Not all genes with the same expression profile share a particular motif. Also, the same transcription factor can be involved in the regulation of a wide variety of expression profiles. In some cases, response may be concentration-dependent so that the same factor can be an activator at some concentrations and a repressor at others. The distance between the binding motif and transcription start site (TSS) may affect the efficiency of transcription. Some transcription factors may require co-factors. Thus regulatory element function is heavily context-dependent (Fessele et al 2002, Herault et al 1999, Werner et al 2003). While it is possible to identify conserved elements, there is often no one-to-one correspondence between the presence of that element and the transcription profile. Importantly, the binding sites that confer important biological responses may often be the ones with lowest affinity for the transcription factor which can be more sensitive to slight changes in protein concentration. Thus the sites that diverge most from the consensus sequence may be very important in regulation. Furthermore, in higher eukaryotes, regulatory elements tend to be dispersed over tens and even hundreds of kilobases, and can be positioned downstream of, within, or upstream of genes. This greatly increases the amount of

space that must be searched in order to find these motifs.

For a group of genes, if we have no prior knowledge about what transcription factors regulate them, the common strategy for identifying motifs in their promoters is Gibbs Sampling (McCue et al 2002, Newberg et al 2007, Thompson et al 2005, Thompson et al 2004, Thompson et al 2003, Thompson et al 2007). The strategy is to iterate randomly through ungapped alignments of the promoter sequences, searching for alignments that result in the identification of blocks of conserved residues of some pre-specified length. In essence, the result is an optional local multiple sequence alignment.

An alternate strategy for understanding the co-regulation of genes in the same cluster or similar functional group (for example Gene Ontology terms) is to identify candidate transcription factors first. The transcription factors can be identified from the microarray data itself by searching for the TFs with highest expression levels. Candidate TFs can also be identified by literature analysis or pathway analysis (Cartharius et al 2005, Quandt et al 1995, Werner 2000).

Once candidate TFs are identified, we can search their binding sequences by mining literature or some TF database, such as TRANSFAC, JASPAR or Genomatix. Then we search their binding sites in promoters of the interesting candidate genes. If there is only one binding site (one exact short sequence), the search will be simple and fast; however, variants will be missed. Since TFBSs are typically degenerate and there are often several variations of binding sites for one transcription factor, a common approach is to build a binding matrix then search this binding matrix (position weight matrices, or PWM) in gene promoter sequences (Stormo 2000a, Stormo 2000b). It is

important to compare such a binding site search in a group of candidate genes to a control group of genes. For TFBS analysis, as is done in this thesis, several control groups can be randomly extracted from the same microarray. The genes in the control groups may not be expressed at all or may be expressed at different levels for different samples or conditions. If specific TFs appear to be enriched in the candidate group of genes compared to the control group(s), the functional basis for this will need to be verified using laboratory studies. It is important to note that, given the correct binding matrix, binding sites will probably bind to the corresponding factors in in vitro studies, such as EMSA. But this does not mean that these sites are necessarily bound in vivo. Only the context can tell a functional site from a mere physical binding site (Elkon et al 2003). TFBS analysis tools such as MatInspector used in this thesis work can identify the sites but cannot determine the functionality of the sites. Most sites are only functional in certain cells, tissues or developmental stages. The biological function of each site can only be proven experimentally. Advanced bioinformatic strategies including orthologous or comparative genomic analysis can reduce the number of the candidate sites that need to be tested.

Interpreting microarray data from an interaction network view

The traditional method of microarray study is to find the list of significantly changed genes and then do further experimental analysis based on this gene list. However, sets of genes act in concert. Therefore, single-gene analysis may miss important effects on pathways. For example, an increase of small percentage (10% or 20%) in all genes of a metabolic pathway may dramatically alter the flux through that pathway and may be more important than a 20-fold up-regulation in a single gene. For this reason, it is important to understand the identified gene list in the context of interaction networks.

Researchers have proposed a couple alternative ways to mine the microarray data. There are several tools to perform a global analysis of gene expression or genomic data in the context of biological pathways and groupings of genes. One of them is GenMAPP (www.genmapp.org). It is a free computer application designed by the Conklin Lab in UCSF to visualize gene expression and other genomic data on maps representing pathways and Gene Ontology Term Sets (Dahlquist 2004, Dahlquist et al 2002, Salomonis et al 2007). The GenMAPP database has hundreds of pathways (signaling pathways, metabolic pathways, physiological pathways) and thousands of Gene Ontology terms which are manually made by scientists based on currently known information. The input to GenMAPP for microarray data analysis is the list of significantly genes or the entire gene list on the microarray with fold change and p-values. GenMAPP can visualize selected pathways by searching the input data based on the user-set criteria. These criteria include fold change, p-value and colors for different input data information to color the genes in the pathway map. Users can even build their own pathway based on the known interaction network. GenMAPP is a powerful tool to visualize gene interaction in a pathway format and can facilitate the understanding of microarray data. However, the disadvantage is that it relies on currently known information about pathways and there is no way for GenMAPP to identify new interaction relationships between genes.

Another tool for understanding gene interactions is GeneGo (www.genego.com) (Ekins 2006, Ekins et al 2006, Ekins et al 2005a, Ekins et al 2007, Ekins et al 2005b, Nikolsky et al 2005a, Nikolsky et al 2005b). GeneGo is a leading bioinformatics software package for data mining applications in microarray data and systems biology. MetaBase is GeneGo's manually-curated database of mammalian biology and

medicinal chemistry. This database is built and constantly curated by entering increasingly published gene interaction information which is obtained by scientists who read the literature every day and identify new gene interaction relationships. The input for GeneGo is a list of genes identified using statistical methods. Often, genes in the input list are differentially expressed in certain tissues, cells or conditions. GeneGo builds the gene-interaction network by searching gene interaction relationship information in its expert-manually-curated database for the input list of genes. GeneGo can also be used for Gene Ontology (GO) or MeSH term enrichment analysis which may reveal important biological themes. GeneGo has incorparated visualization power by using different symbols for different types of genes; for example, transcription factors are represented by one special symbol while kinases by another. Furthermore, the interaction networks built by GeneGo are dynamic because each connection line between two genes is a link to the literature in which the interaction information is obtained. Researchers can view this information by clicking the connection line between the genes in the network map. This is totally different from the GenMAPP pathway maps which are static and only the fold change or gene expression value and/or p-value are shown on the map. GeneGo can also build transcription regulation networks based on the published regulation information but cannot find new transcriptional regulation networks between genes. GeneGo can also filter longer lists based on gene expression level, fold change or p-value before building the interaction networks. The disadvantage of GeneGo is that it performs best for short lists (less than 50 or so). If the input gene list is too long, for example, several hundred genes, which is normal in microarray analysis, GeneGo will build very complex networks; this can be hard to interpret. However, the interaction relationships in the expert-manually-curated database are more reliable than those in

other databases that use computer natural language processing software to find relationships.

Another tool based on literature mining is the Genomatix BiblioSphere Pathway Edition (www.genomatix.de) (Cartharius et al 2005, Fessele et al 2002, Klingenhoff et al 2002, Scherf et al 2005, Seifert et al 2005, Werner 2000, Werner 2001a, Werner 2001b, Werner 2002a, Werner 2002b, Werner 2007, Werner 2008, Werner et al 2003, Werner & Nelson 2006). Like GeneGo, the gene interaction information in BiblioSphere is also based on literature which is curated by experts and processed by automatic computer natural language processing algorithms. BiblioSphere not only uses gene interaction but also includes gene co-citation information in its database to filter the input list. Like GeneGo, BiblioSphere can also do GO term analysis and follow-up pathway analysis; the network maps are also dynamic to show literature information. The big differences between GeneGo and BiblioSphere are that BiblioSphere integrates transcription factor binding site analysis in its network building process and that it can also show pathway information (signaling, metabolic) on the networks. The gene network analysis, co-citation analysis, GO term analysis and transcription regulation analysis are well integrated in BiblioSphere. The advantage is that the results from BiblioSphere analysis can be easily followed up by other Genomatix tools for TFBS framework analysis, comparative genomic analysis, etc. The disadvantage is that the analysis following up BiblioSphere analysis, especially for TFBS framework or module analysis, only works best for very short list (about 10 or 20 genes).

Both GeneGo and Genomatix are designed to ideally perform well on a list of genes that are identified by previous statistical tests. However, for microarray experiments

in which few genes changed their expression across the conditions or samples, we may end up with a list with only a small number of genes or even no significant genes found after the consideration of multiple hypotheses testing. Or in some other experiments in which many genes changed expression levels, we may have a long list which contains thousands of genes and it is hard to tell the main biological theme for these genes in the list. Neither of these situations is very good for pathway or interaction analysis discussed above.

Another problem is that, since we usually select significantly changed genes based on p-value and fold change (FC), there are some genes which may not meet the cutoff threshold due to the small sample size (which usually happens in microarray experiments due to the high cost or sample availability) but these genes may have important biological functions. For example, some critical transcriptional factors or signal modulators may be expressed at very low levels or in only a few cells. Another problem arising from low-expressing genes is that when different groups study the same biological model, the list of statistically significant genes from the two groups may show little overlap. In order to overcome these challenges, a method called Gene Set Enrichment Analysis (GSEA) was developed to interpret microarray data without identifying the gene list first and evaluate the data at the level of predefined gene sets (http://www.broad.mit.edu/gsea/) (Mootha et al 2003, Subramanian et al 2005). This algorithm uses the genome-wide expression profiles in the microarray experiment as input, not just a list of significantly changed genes; this is completely different from GeneGo or Genomatix (BiblioSphere). Genes are ranked first according to their differential expression between the classes. Except for using the entire gene expression profiles, this method derives its power by focusing on gene sets, groups of

genes that share common biological function, chromosomal location, or regulation. The GSEA database, Molecular Signatures Database (MSigDB), contains more than 5000 gene sets for use with GSEA (www.broad.mit.edu/gsea/msigdb/). The MSigDB gene sets are divided into five major collections: C1, positional gene sets for each human chromosome and each cytogenetic band; C2, curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts; C3, motif gene sets based on conserved *cis*-regulatory motifs from a comparative analysis of the human, mouse, rat and dog genomes; C4, computational gene sets defined by expression neighborhoods centered on 380 cancer-associated genes, and C5, GO gene sets annotated by the same GO terms. Users are also allowed to define their own gene sets. The goal of GSEA is to determine whether members in a pre-defined gene set are randomly distributed throughout the entire ranked gene list (in which case the gene set is not correlated with the compared two conditions and is not considered significant), or primarily found at the top or bottom of the ranked list (in which case the gene set is correlated with the phenotypic class distinction). GSEA first calculates an enrichment score (ES) that reflects the degree to which a given set is overrepresented at the top or bottom of the entire ranked list. Then, it estimates the statistical significance of the ES by using a permutation test procedure. Finally, it adjusts the estimated significance level to account for multiple hypothesis testing by FDR control. The power of GSEA is that it evaluates microarray data in gene sets and using the whole gene expression profiles without losing low-level expression genes. However, it is still based on the known information and cannot discover new interactions.

The methods discussed above help to interpret microarray data by combining the gene expression information from microarray experiments with other biological

information (literature mining, pathways, GO terms analysis, TFBS analysis, etc). Other alternate ways to evaluate microarray data could be integration of protein-protein interaction data, other microarray data (GEO database), CHIP-on-Chip data, SAGE, etc.

Verification and follow-up experiments for microarray results

Like other high-throughput and semi-quantitative methods, noise is unavoidably introduced in the microarray experiment process and there is a risk of making false calls. Furthermore, there are possibilities of making errors in the probe selection and probe design (there are several thousands probesets not assigned with gene annotation during cross-gene-hybridization). In addition, microarray technology detects mRNA level changes and mRNAs are unstable biomolecules. The changes in mRNA level do not necessarily indicate protein level changes because important regulations take place at the level of translation and protein modification. For these reasons, one of the tasks in the further analysis following microarray experiments is to verify important findings using different experimental procedures. Several traditional single-gene analysis methods are available, including QRT-PCR, *in situ*, Northern blotting, immunohistochemistry, etc. QRT-PCR is quantitative real time PCR for measuring mRNA levels; it is the most rapid method of verification, especially for a large number of genes. Northern blotting is an alternative way to investigate mRNA levels, but it requires much more template RNA than PCR does. *in situ* hybridization or immunohistochemistry are effective ways to determine gene expression pattern in vivo ; these methods are very attractive when dealing with heterogeneous tissues as in this thesis because they reveal the spatial and temporal expression pattern of the genes. The interesting genes or the generated hypothesis can also be functionally studied by

knock out or transgenic mouse models or in tissue culture analyses. The functional studies are the most time consuming, but of course are also the most important.

Summary

In this thesis, microarray analysis and bioinformatics tools were used to probe basic questions in gut development. In chapter two, I present the results of an examination of mouse intestinal epithelial versus mesenchymal gene expression profiles. This analysis is important because it allows us to begin to understand how epithelial and mesenchymal cells produce and receive signals that regulate patterning and homeostasis of the small intestine. In chapter three, I investigate a basic patterning question: what genes are changed during the establishment of intestinal cell identity during pyloric border formation? This work establishes for the first time, that generation of a distinct epithelial pyloric border involves the coordinate up-regulation of over a thousand genes in intestinal (not stomach) epithelium. We also identified genes that are specifically regulated right at the pyloric border itself and might be responsible for border establishment. This study has implications for gut patterning, intestinal cell differentiation and for abnormal lesions such as intestinal metaplasia. Finally, in chapter four, I summarize these results in light of other studies and identify some of the most important future questions for study.

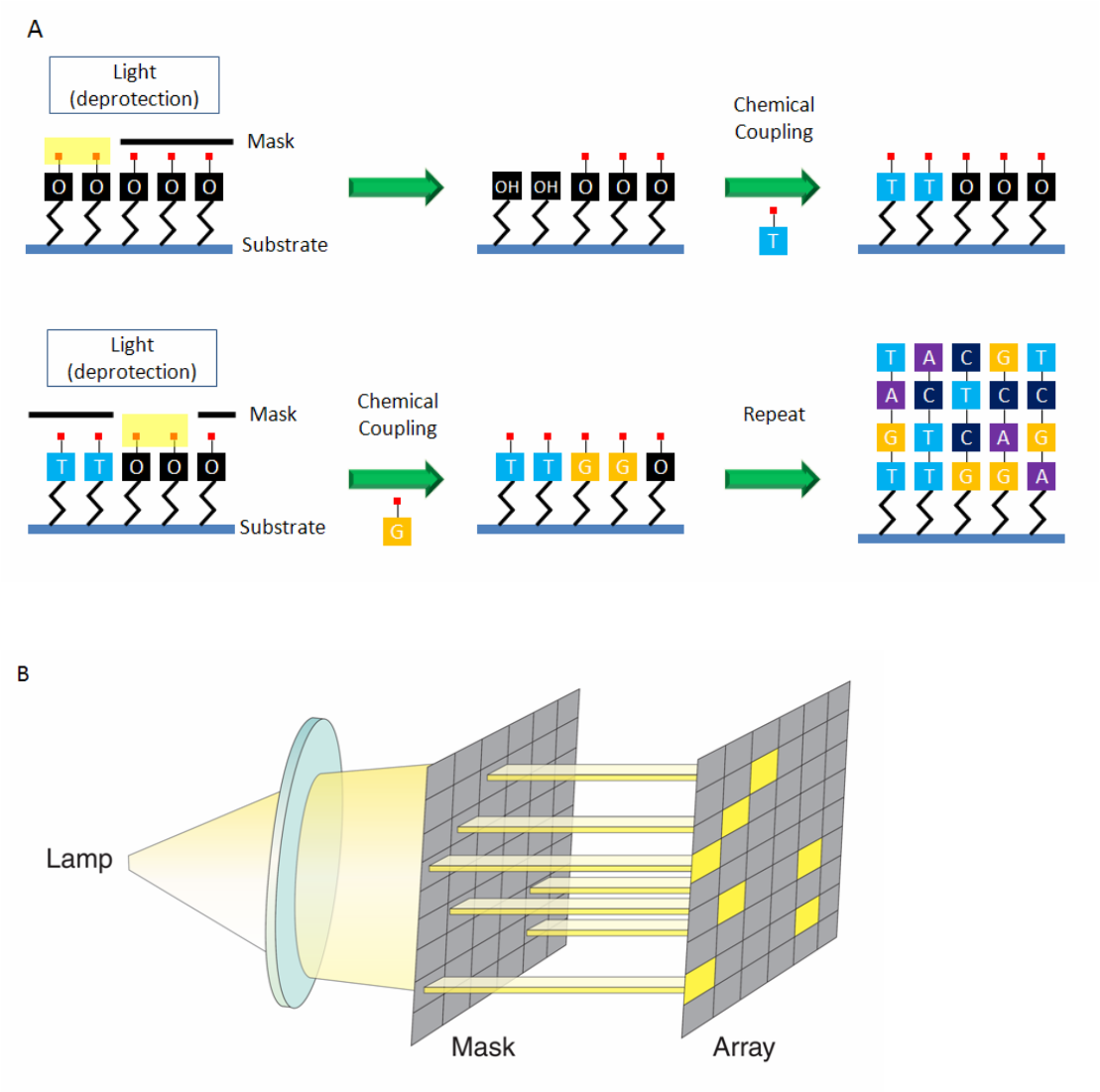

Figure 1.1 Manufacture of Affymetrix high-density oligonucleotide arrays. Affymetrix GeneChips are constructed by photolithography and solid-phase oligonucleotide synthesis. A. light-directed oligonucleotide probe synthesis. A solid substrate is derivatized with a covalent linker molecule ended with a photolabile protecting group. Light is directed through a mask to deprotect and activate the selected area, and protected nucleotides couple to the activated sites. This process is repeated by activating different sets of sites and coupling different bases allowing arbitrary DNA probes to be synthesized at each base. B. Schematic representation of the lamp, mask and array (Image courtesy of Affymetrix).

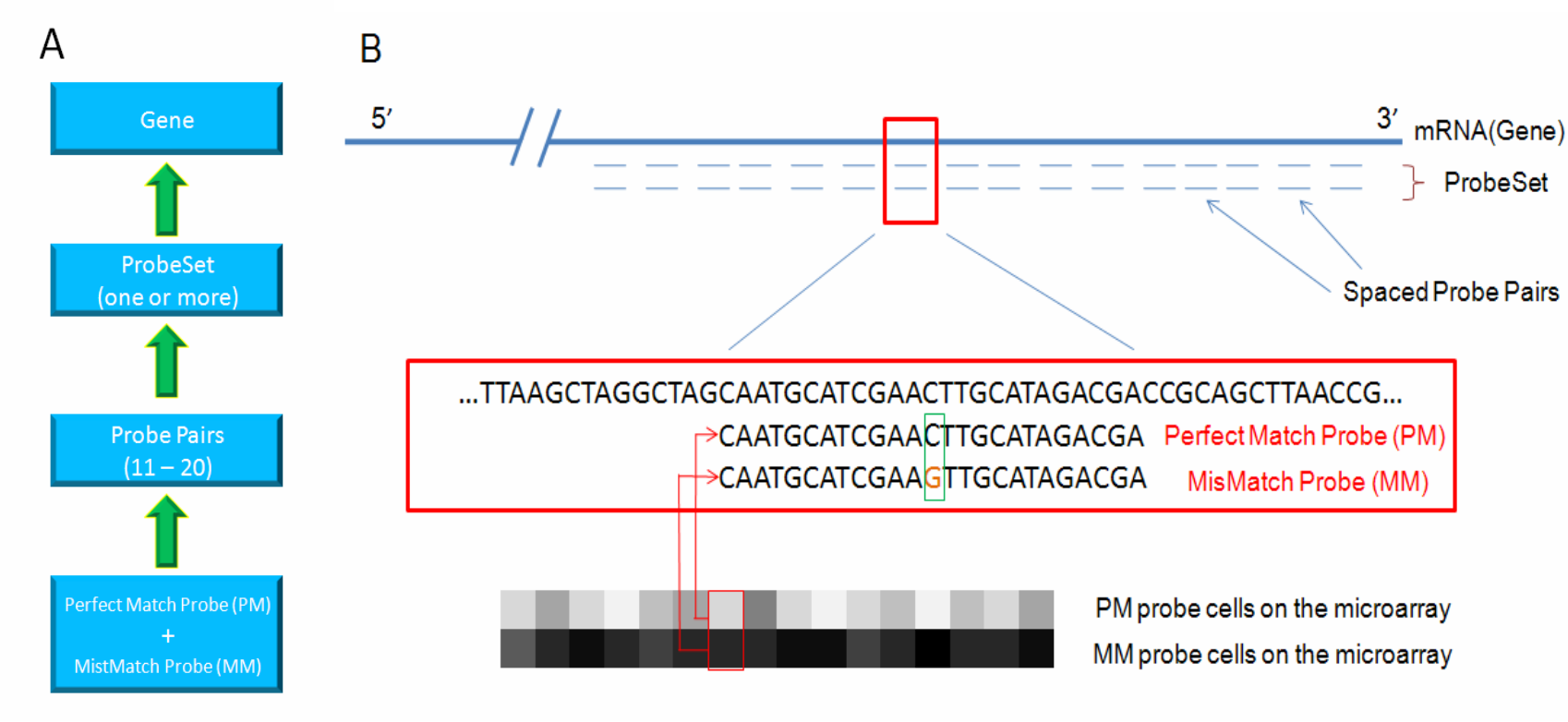

Figure 1.2 The Affymetrix GeneChip microarray: From probes to gene. The relative amount of mRNA in samples (tissues or cells) is detected by one or more ProbeSets which consist of a set of probe pairs (16-20). Each probe pair is composed of Perfect Match (PM) and MisMatch (MM), length of 25 bases. Note the base at the middle position (13^{th}) of MM is changed compared to PM (indicated by green rectangle). Hybridization of labeled samples on a GeneChip is detected by laser scanning of the chip. In the schematic, the level of gray of black color shows different amounts of hybridization.

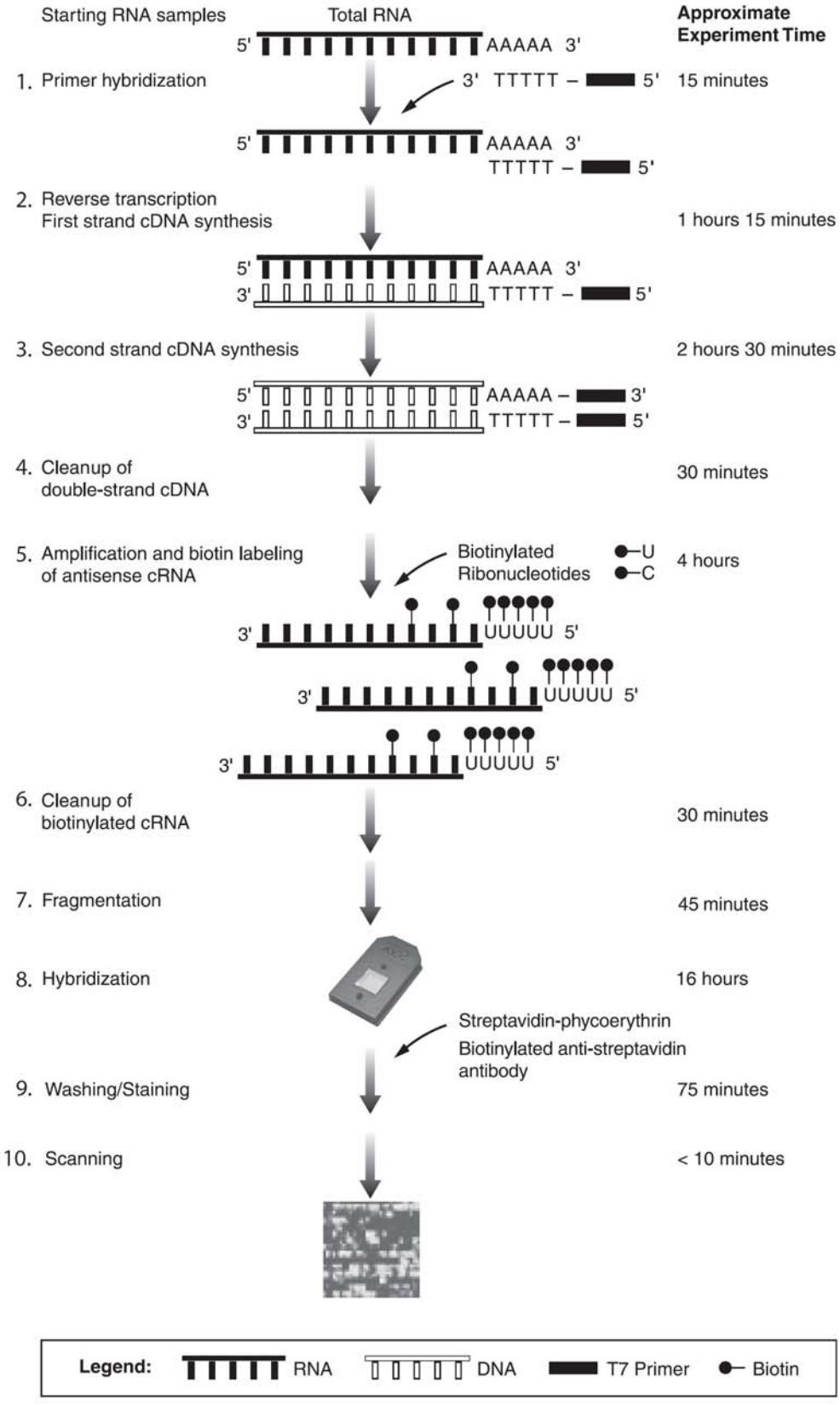

Figure 1.3 The process of a microarray experiment using the Affymetrix GeneChip for eukaryotic study (Image courtesy of Affymetrix).

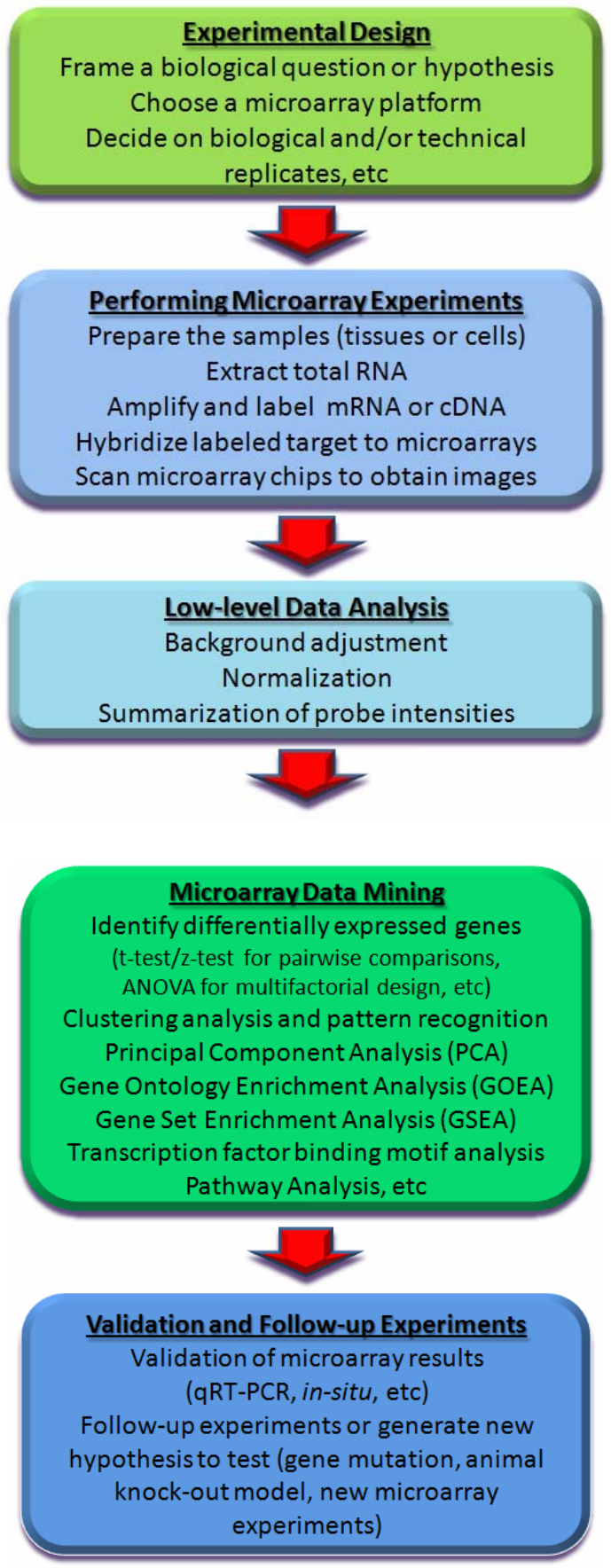

Figure 1.4 Flowchart of microarray experiment analysis

Table 1.1 Summary of methods for microarray data analysis.

Methods	Background Adjustment	Normalization	Usage of MisMatch	Summarization of Probeset Intensities	Literature
MAS4.0	Reginal Adjustment	scaling by a constant	MisMatch	AveDiff	(Affymetrix 1999)
MAS5.0	Reginal Adjustment	scaling by a constant	ideal MisMatch	Tukey Biweight Average	(Affymetrix 2001, Affymetrix 2002)
RMA	whole array adjustment	quantile	No	medianpolish	(Bolstad et al 2003, Irizarry et al 2003a, Irizarry et al 2003b)
GCRMA	GC content of probes	quantile	Yes/No	medianpolish	(Wu & Irizarry 2004, Wu & Irizarry 2005)
dChip	None	invariant set	Yes /No	MBEI (Li-Wong multiplicative model)	(Li & Hung Wong 2001, Li & Wong 2001)

Bibliography

Affymetrix. 1999. Affymetrix microarray suite users guide. *Santa clara, CA* Affymetrix, 1999

Affymetrix. 2001. Affymetrix microarray suite users guide, Version 5.0 ed. *Santa clara, CA* Affymetrix, 2001

Affymetrix. 2002. Statistical algorithms description document *Santa clara, CA* Affymetrix

Ahlgren U, Jonsson J, Edlund H. 1996. The morphogenesis of the pancreatic mesenchyme is uncoupled from that of the pancreatic epithelium in IPF1/PDX1-deficient mice. *Development* 122: 1409-16

Ahlgren U, Pfaff SL, Jessell TM, Edlund T, Edlund H. 1997. Independent requirement for ISL1 in formation of pancreatic mesenchyme and islet cells. *Nature* 385: 257-60

Alter O, Brown PO, Botstein D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* 97: 10101-6

Alter O, Brown PO, Botstein D. 2003. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc Natl Acad Sci U S A* 100: 3351-6

Andreu P, Peignon G, Slomianny C, Taketo MM, Colnot S, et al. 2008. A genetic study of the role of the Wnt/beta-catenin signalling in Paneth cell differentiation. *Dev Biol*

Ang SL, Jin O, Rhinn M, Daigle N, Stevenson L, Rossant J. 1996. A targeted mouse Otx2 mutation leads to severe defects in gastrulation and formation of axial mesoderm and to deletion of rostral brain. *Development* 122: 243-52

Artavanis-Tsakonas S, Rand MD, Lake RJ. 1999. Notch signaling: cell fate control and signal integration in development. *Science* 284: 770-6

Astrand M. 2003. Contrast normalization of oligonucleotide arrays. *J Comput Biol* 10: 95-102

Auclair BA, Benoit YD, Rivard N, Mishina Y, Perreault N. 2007. Bone morphogenetic protein signaling is essential for terminal differentiation of the intestinal secretory cell lineage. *Gastroenterology* 133: 887-96

Baldi P, Long AD. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* 17: 509-19

Baron M. 2003. An overview of the Notch signalling pathway. *Semin Cell Dev Biol* 14: 113-9

Batlle E, Henderson JT, Beghtel H, van den Born MM, Sancho E, et al. 2002. Beta-catenin and TCF mediate cell positioning in the intestinal epithelium by controlling the expression of EphB/ephrinB. *Cell* 111: 251-63

Beck F. 2002. Homeobox genes in gut development. *Gut* 51: 450-4

Beck F. 2004. The role of Cdx genes in the mammalian gut. *Gut* 53: 1394-6

Beck F, Erler T, Russell A, James R. 1995. Expression of Cdx-2 in the mouse embryo and placenta: possible role in patterning of the extra-embryonic membranes. *Dev Dyn* 204: 219-27

Beck F, Tata F, Chawengsaksophak K. 2000. Homeobox genes and gut development. *Bioessays* 22: 431-41

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc* B 57: 289-300

Biben C, Hatzistavrou T, Harvey RP. 1998a. Expression of NK-2 class homeobox gene Nkx2-6 in foregut endoderm and heart. *Mech Dev* 73: 125-7

Biben C, Stanley E, Fabri L, Kotecha S, Rhinn M, et al. 1998b. Murine cerberus homologue mCer-1: a candidate anterior patterning molecule. *Dev Biol* 194: 135-51

Bienz M, Clevers H. 2000. Linking colorectal cancer to Wnt signaling. *Cell* 103: 311-20

Bitgood MJ, Shen L, McMahon AP. 1996. Sertoli cell signaling by Desert hedgehog regulates the male germline. *Curr Biol* 6: 298-304

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

Booth C, Brady G, Potten CS. 2002. Crowd control in the crypt. *Nat Med* 8: 1360-1

Bort R, Signore M, Tremblay K, Martinez Barbera JP, Zaret KS. 2006. Hex homeobox gene controls the transition of the endoderm to a pseudostratified, cell emergent epithelium for liver bud development. *Dev Biol* 290: 44-56

Boucher MJ, Simoneau M, Edlund H. 2008. The Homeodomain-interacting protein kinase 2(HIPK2) regulates IPF1/PDX1 transcriptional activity. *Endocrinology*

Braunstein EM, Qiao XT, Madison B, Pinson K, Dunbar L, Gumucio DL. 2002. Villin: A marker for development of the epithelial pyloric border. *Dev Dyn* 224: 90-102

Calvert R, Pothier P. 1990. Migration of fetal intestinal intervillous cells in neonatal mice. *Anat Rec* 227: 199-206

Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, et al. 2005. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics* 21: 2933-42

Choi MY, Romer AI, Hu M, Lepourcelet M, Mechoor A, et al. 2006. A dynamic expression survey identifies transcription factors relevant in mouse digestive tract development. *Development* 133: 4119-29

Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, et al. 1998. The transcriptional program of sporulation in budding yeast. *Science* 282: 699-705

Chu TM, Weir B, Wolfinger R. 2002. A systematic statistical linear modeling approach to oligonucleotide array experiments. *Math Biosci* 176: 35-51

Clarke AR. 2006. Wnt signalling in the mouse intestine. *Oncogene* 25: 7512-21

Clevers H. 2006. Wnt/beta-catenin signaling in development and disease. *Cell* 127: 469-80

Dahlquist KD. 2004. Using GenMAPP and MAPPFinder to view microarray data on biological pathways and identify global trends in the data. *Curr Protoc Bioinformatics* Chapter 7: Unit 7 5

Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR. 2002. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* 31: 19-20

de Lau W, Barker N, Clevers H. 2007. WNT signaling in the normal intestine and colorectal cancer. *Front Biosci* 12: 471-91

Desai S, Loomis Z, Pugh-Bernard A, Schrunk J, Doyle MJ, et al. 2008. Nkx2.2 regulates cell fate choice in the enteroendocrine cell lineages of the intestine. *Dev Biol* 313: 58-66

Do JH, Choi DK. 2008. Clustering approaches to identifying gene expression patterns from DNA microarray data. *Mol Cells* 25: 279-88

Doyle MJ, Sussel L. 2007. Nkx2.2 regulates beta-cell function in the mature islet. *Diabetes* 56: 1999-2007

Dudoit S, Yang YH, Callow MJ, Speed TP. 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 12: 111-39

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

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

Ekins S. 2006. Systems-ADME/Tox: resources and network approaches. *J Pharmacol Toxicol Methods* 53: 38-66

Ekins S, Bugrim A, Brovold L, Kirillov E, Nikolsky Y, et al. 2006. Algorithms for network analysis in systems-ADME/Tox using the MetaCore and MetaDrug platforms. *Xenobiotica* 36: 877-901

Ekins S, Kirillov E, Rakhmatulin EA, Nikolskaya T. 2005a. A novel method for visualizing nuclear hormone receptor networks relevant to drug metabolism. *Drug Metab Dispos* 33: 474-81

Ekins S, Nikolsky Y, Bugrim A, Kirillov E, Nikolskaya T. 2007. Pathway mapping tools for analysis of high content data. *Methods Mol Biol* 356: 319-50

Ekins S, Nikolsky Y, Nikolskaya T. 2005b. Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity. *Trends Pharmacol Sci* 26: 202-9

Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y. 2003. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res* 13: 773-80

Epstein M, Pillemer G, Yelin R, Yisraeli JK, Fainsod A. 1997. Patterning of the embryo along the anterior-posterior axis: the role of the caudal genes. *Development* 124: 3805-14

Fang R, Olds LC, Sibley E. 2006. Spatio-temporal patterns of intestine-specific transcription factor expression during postnatal mouse gut development. *Gene Expr Patterns* 6: 426-32

Farzan SF, Singh S, Schilling NS, Robbins DJ. 2008. The adventures of sonic hedgehog in development and repair. III. Hedgehog processing and biological activity. *Am J Physiol Gastrointest Liver Physiol* 294: G844-9

Fessele S, Maier H, Zischek C, Nelson PJ, Werner T. 2002. Regulatory context is a crucial part of gene function. *Trends Genet* 18: 60-3

Fukuda K, Yasugi S. 2002. Versatile roles for sonic hedgehog in gut development. *J Gastroenterol* 37: 239-46

Giles PJ, Kipling D. 2003. Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics* 19: 2254-62

Green RP, Cohn SM, Sacchettini JC, Jackson KE, Gordon JI. 1992. The mouse intestinal fatty acid binding protein gene: nucleotide sequence, pattern of developmental and regional expression, and proposed structure of its protein product. *DNA Cell Biol* 11: 31-41

Gregorieff A, Clevers H. 2005. Wnt signaling in the intestinal epithelium: from endoderm to cancer. *Genes Dev* 19: 877-90

Gregorieff A, Pinto D, Begthel H, Destree O, Kielman M, Clevers H. 2005. Expression pattern of Wnt signaling components in the adult intestine. *Gastroenterology* 129: 626-38

Hansson EM, Lendahl U, Chapman G. 2004. Notch signaling in development and disease. *Semin Cancer Biol* 14: 320-8

Haramis AP, Begthel H, van den Born M, van Es J, Jonkheer S, et al. 2004. De novo crypt formation and juvenile polyposis on BMP inhibition in mouse intestine. *Science* 303: 1684-6

Hardwick JC, Van Den Brink GR, Bleuming SA, Ballester I, Van Den Brande JM, et al. 2004. Bone morphogenetic protein 2 is expressed by, and acts upon, mature epithelial cells in the colon. *Gastroenterology* 126: 111-21

Herault Y, Beckers J, Gerard M, Duboule D. 1999. Hox gene expression in limbs: colinearity by opposite regulatory controls. *Dev Biol* 208: 157-65

Hocker M, Wiedenmann B. 1998. Molecular mechanisms of enteroendocrine differentiation. *Ann N Y Acad Sci* 859: 160-74

Hoffmann R, Seidl T, Dugas M. 2002. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol* 3: RESEARCH0033

Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV. 2000. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc Natl Acad Sci U S A* 97: 8409-14

Howe JR, Bair JL, Sayed MG, Anderson ME, Mitros FA, et al. 2001. Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis. *Nat Genet* 28: 184-7

Howe JR, Roth S, Ringold JC, Summers RW, Jarvinen HJ, et al. 1998. Mutations in the SMAD4/DPC4 gene in juvenile polyposis. *Science* 280: 1086-8

Huang D, Chen SW, Langston AW, Gudas LJ. 1998. A conserved retinoic acid responsive element in the murine Hoxb-1 gene is required for expression in the developing gut. *Development* 125: 3235-46

Hubbell E, Liu WM, Mei R. 2002. Robust estimators for expression analysis. *Bioinformatics* 18: 1585-92

Ingham PW, McMahon AP. 2001. Hedgehog signaling in animal development: paradigms and principles. *Genes Dev* 15: 3059-87

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. 2003a. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31: e15

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. 2003b. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4: 249-64

Ishizuya-Oka A, Hasebe T. 2008. Sonic hedgehog and bone morphogenetic protein-4 signaling pathway involved in epithelial cell renewal along the radial axis of the intestine. *Digestion* 77 Suppl 1: 42-7

Ishizuya-Oka A, Hasebe T, Shimizu K, Suzuki K, Ueda S. 2006. Shh/BMP-4 signaling pathway is essential for intestinal epithelial development during Xenopus larval-to-adult remodeling. *Dev Dyn* 235: 3240-9

Iso T, Kedes L, Hamamori Y. 2003. HES and HERP families: multiple effectors of the Notch signaling pathway. *J Cell Physiol* 194: 237-55

Jensen J, Pedersen EE, Galante P, Hald J, Heller RS, et al. 2000. Control of endodermal endocrine development by Hes-1. *Nat Genet* 24: 36-44

Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. 2001. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nat Genet* 29: 389-95

Kedinger M, Lefebvre O, Duluc I, Freund JN, Simon-Assmann P. 1998. Cellular and molecular partners involved in gut morphogenesis and differentiation. *Philos Trans R Soc Lond B Biol Sci* 353: 847-56

Kerr MK, Churchill GA. 2007. Statistical design and the analysis of gene expression microarray data. *Genet Res* 89: 509-14

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

Kim BM, Miletich I, Mao J, McMahon AP, Sharpe PA, Shivdasani RA. 2007. Independent functions and mechanisms for homeobox gene Barx1 in patterning mouse stomach and spleen. *Development* 134: 3603-13

Kinzler KW, Vogelstein B. 1996. Lessons from hereditary colorectal cancer. *Cell* 87: 159-70

Klingenhoff A, Frech K, Werner T. 2002. Regulatory modules shared within gene classes as well as across gene classes can be detected by the same in silico approach. *In Silico Biol* 2: S17-26

Kohler EM, Vijaya Chandra SH, Behrens J, Schneikert J. 2008. {beta}-catenin degradation mediated by the CID domain of APC provides a model for the selection of APC mutations in colorectal, desmoid and duodenal tumours. *Hum Mol Genet*

Korinek V, Barker N, Moerer P, van Donselaar E, Huls G, et al. 1998. Depletion of epithelial stem-cell compartments in the small intestine of mice lacking Tcf-4. *Nat Genet* 19: 379-83

Larsson LI, St-Onge L, Hougaard DM, Sosa-Pineda B, Gruss P. 1998. Pax 4 and 6 regulate gastrointestinal endocrine cell development. *Mech Dev* 79: 153-9

Lawson KA, Pedersen RA. 1987. Cell fate, morphogenetic movement and population kinetics of embryonic endoderm at the time of germ layer formation in the mouse. *Development* 101: 627-52

Lees C, Howie S, Sartor RB, Satsangi J. 2005. The hedgehog signalling pathway in the gastrointestinal tract: implications for development, homeostasis, and disease. *Gastroenterology* 129: 1696-710

Lemon WJ, Palatini JJ, Krahe R, Wright FA. 2002. Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics* 18: 1470-6

Li C, Hung Wong W. 2001. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* 2: RESEARCH0032

Li C, Wong WH. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A* 98: 31-6

Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. 1999. High density synthetic oligonucleotide arrays. *Nat Genet* 21: 20-4

Liu YZ, Xia ZX, Liu XL, Huang Q. 2003. [Expression of Pax-6 homeobox gene in lens epithelial cells in vitro]. *Zhonghua Yan Ke Za Zhi* 39: 395-9

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 14: 1675-80

Lonnstedt I, Speed TP. 2002. Replicated Microarray Data. *Statistica Sinica* 12: 31-46

Ma J, Qin ZS. 2007. Different normalization strategies for microarray gene expression traits affect the heritability estimation. *BMC Proc* 1 Suppl 1: S154

Madison BB, Braunstein K, Kuizon E, Portman K, Qiao XT, Gumucio DL. 2005. Epithelial hedgehog signals pattern the intestinal crypt-villus axis. *Development* 132: 279-89

Madison BB, Dunbar L, Qiao XT, Braunstein K, Braunstein E, Gumucio DL. 2002. Cis elements of the villin gene control expression in restricted domains of the vertical (crypt) and horizontal (duodenum, cecum) axes of the intestine. *J Biol Chem* 277: 33275-83

Maeda Y, Hunter TC, Loudy DE, Dave V, Schreiber V, Whitsett JA. 2006. PARP-2 interacts with TTF-1 and regulates expression of surfactant protein-B. *J Biol Chem* 281: 9600-6

Mansouri A, Chowdhury K, Gruss P. 1998. Follicular cells of the thyroid gland require Pax8 gene function. *Nat Genet* 19: 87-90

Martin CC, Oeser JK, O'Brien RM. 2004. Differential regulation of islet-specific glucose-6-phosphatase catalytic subunit-related protein gene transcription by Pax-6 and Pdx-1. *J Biol Chem* 279: 34277-89

Mathan M, Moxey PC, Trier JS. 1976. Morphogenesis of fetal rat duodenal villi. *Am J Anat* 146: 73-92

McCue LA, Thompson W, Carmack CS, Lawrence CE. 2002. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res* 12: 1523-32

Minoo P, Hamdan H, Bu D, Warburton D, Stepanik P, deLemos R. 1995. TTF-1 regulates lung epithelial morphogenesis. *Dev Biol* 172: 694-8

Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, et al. 2003. Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc Natl Acad Sci U S A* 100: 605-10

Mumm JS, Kopan R. 2000. Notch signaling: from the outside in. *Dev Biol* 228: 151-65

Naef F, Lim DA, Patil N, Magnasco M. 2002. DNA hybridization to mismatched templates: a chip study. *Phys Rev E Stat Nonlin Soft Matter Phys* 65: 040902

Newberg LA, Thompson WA, Conlan S, Smith TM, McCue LA, Lawrence CE. 2007. A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction. *Bioinformatics* 23: 1718-27

Nielsen CM, Williams J, van den Brink GR, Lauwers GY, Roberts DJ. 2004. Hh pathway expression in human gut tissues and in inflammatory gut diseases. *Lab Invest* 84: 1631-42

Nikolova M, Chen X, Lufkin T. 1997. Nkx2.6 expression is transiently and specifically restricted to the branchial region of pharyngeal-stage mouse embryos. *Mech Dev* 69: 215-8

Nikolsky Y, Ekins S, Nikolskaya T, Bugrim A. 2005a. A novel method for generation of signature networks as biomarkers from complex high throughput data. *Toxicol Lett* 158: 20-9

Nikolsky Y, Nikolskaya T, Bugrim A. 2005b. Biological networks and analysis of experimental data in drug discovery. *Drug Discov Today* 10: 653-62

Nybakken K, Perrimon N. 2002. Hedgehog signal transduction: recent findings. *Curr Opin Genet Dev* 12: 503-11

Omenetti A, Diehl AM. 2008. The adventures of sonic hedgehog in development and repair. II. Sonic hedgehog and liver development, inflammation, and cancer. *Am J Physiol Gastrointest Liver Physiol* 294: G595-8

Parkin CA, Ingham PW. 2008. The adventures of Sonic Hedgehog in development and repair. I. Hedgehog signaling in gastrointestinal development and disease. *Am J Physiol Gastrointest Liver Physiol* 294: G363-7

Parrish RS, Spencer HJ, 3rd. 2004. Effect of normalization on significance testing for oligonucleotide microarrays. *J Biopharm Stat* 14: 575-89

Perea-Gomez A, Lawson KA, Rhinn M, Zakin L, Brulet P, et al. 2001. Otx2 is required for visceral endoderm movement and for the restriction of posterior signals in the epiblast of the mouse embryo. *Development* 128: 753-65

Peters H, Neubuser A, Kratochwil K, Balling R. 1998. Pax9-deficient mice lack pharyngeal pouch derivatives and teeth and exhibit craniofacial and limb abnormalities. *Genes Dev* 12: 2735-47

Pinto D, Clevers H. 2005. Wnt control of stem cells and differentiation in the intestinal epithelium. *Exp Cell Res* 306: 357-63

Pinto D, Gregorieff A, Begthel H, Clevers H. 2003. Canonical Wnt signals are essential for homeostasis of the intestinal epithelium. *Genes Dev* 17: 1709-13

Playford RJ. 2002. Homeobox genes: going for growth. *Gut* 50: 447-8

Porter EM, Bevins CL, Ghosh D, Ganz T. 2002. The multifaceted Paneth cell. *Cell Mol Life Sci* 59: 156-70

Qin ZS. 2006. Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics* 22: 1988-97

Qin ZS, McCue LA, Thompson W, Mayerhofer L, Lawrence CE, Liu JS. 2003. Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol* 21: 435-9

Quandt K, Frech K, Karas H, Wingender E, Werner T. 1995. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* 23: 4878-84

Ramalho-Santos M, Melton DA, McMahon AP. 2000. Hedgehog signals regulate multiple aspects of gastrointestinal development. *Development* 127: 2763-72

Raychaudhuri S, Stuart JM, Altman RB. 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput*: 455-66

Reiner A, Yekutieli D, Benjamini Y. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19: 368-75

Reya T, Clevers H. 2005. Wnt signalling in stem cells and cancer. *Nature* 434: 843-50

Reynolds PR, Mucenski ML, Whitsett JA. 2003. Thyroid transcription factor (TTF) -1 regulates the expression of midkine (MK) during lung morphogenesis. *Dev Dyn* 227: 227-37

Rhinn M, Dierich A, Shawlot W, Behringer RR, Le Meur M, Ang SL. 1998. Sequential roles for Otx2 in visceral endoderm and neuroectoderm for forebrain and midbrain induction and specification. *Development* 125: 845-56

Ritz-Laser B, Estreicher A, Gauthier BR, Mamin A, Edlund H, Philippe J. 2002. The pancreatic beta-cell-specific transcription factor Pax-4 inhibits glucagon gene expression through Pax-6. *Diabetologia* 45: 97-107

Roberts DJ. 2000. Molecular mechanisms of development of the gastrointestinal tract. *Dev Dyn* 219: 109-20

Roberts DJ, Johnson RL, Burke AC, Nelson CE, Morgan BA, Tabin C. 1995. Sonic hedgehog is an endodermal signal inducing Bmp-4 and Hox genes during induction and regionalization of the chick hindgut. *Development* 121: 3163-74

Rossi DL, Acebron A, Santisteban P. 1995. Function of the homeo and paired domain proteins TTF-1 and Pax-8 in thyroid cell proliferation. *J Biol Chem* 270: 23139-42

Salomonis N, Hanspers K, Zambon AC, Vranizan K, Lawlor SC, et al. 2007. GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics* 8: 217

Sancho E, Batlle E, Clevers H. 2004. Signaling pathways in intestinal development and cancer. *Annu Rev Cell Dev Biol* 20: 695-723

Sayed MG, Ahmed AF, Ringold JR, Anderson ME, Bair JL, et al. 2002. Germline SMAD4 or BMPR1A mutations and phenotype of juvenile polyposis. *Ann Surg Oncol* 9: 901-6

Schadt EE, Li C, Ellis B, Wong WH. 2001. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J Cell Biochem Suppl* Suppl 37: 120-5

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

Scherf M, Epple A, Werner T. 2005. The next generation of literature analysis: integration of genomic analysis into text mining. *Brief Bioinform* 6: 287-97

Schmidt PH, Lee JR, Joshi V, Playford RJ, Poulsom R, et al. 1999. Identification of a metaplastic cell lineage associated with human gastric adenocarcinoma. *Lab Invest* 79: 639-46

Schroder N, Gossler A. 2002. Expression of Notch pathway components in fetal and adult mouse small intestine. *Gene Expr Patterns* 2: 247-50

Seifert M, Scherf M, Epple A, Werner T. 2005. Multievidence microarray mining. *Trends Genet* 21: 553-8

Sellin JH, Wang Y, Singh P, Umar S. 2008. beta-Catenin stabilization imparts crypt progenitor phenotype to hyperproliferating colonic epithelia. *Exp Cell Res*

Shedden K, Chen W, Kuick R, Ghosh D, Macdonald J, et al. 2005. Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data. *BMC Bioinformatics* 6: 26

Silberg DG, Sullivan J, Kang E, Swain GP, Moffett J, et al. 2002. Cdx2 ectopic expression induces gastric intestinal metaplasia in transgenic mice. *Gastroenterology* 122: 689-96

Smyth GK, Michaud J, Scott HS. 2005. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* 21: 2067-75

Sosa-Pineda B, Chowdhury K, Torres M, Oliver G, Gruss P. 1997. The Pax4 gene is essential for differentiation of insulin-producing beta cells in the mammalian pancreas. *Nature* 386: 399-402

St-Onge L, Sosa-Pineda B, Chowdhury K, Mansouri A, Gruss P. 1997. Pax6 is required for differentiation of glucagon-producing alpha-cells in mouse pancreas. *Nature* 387: 406-9

Stormo GD. 2000a. DNA binding sites: representation and discovery. *Bioinformatics* 16: 16-23

Stormo GD. 2000b. Identification of coordinated gene expression and regulatory sequences. *Pac Symp Biocomput*: 416-7

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102: 15545-50

Sukegawa A, Narita T, Kameda T, Saitoh K, Nohno T, et al. 2000. The concentric structure of the developing gut is regulated by Sonic hedgehog derived from endodermal epithelium. *Development* 127: 1971-80

Sussel L, Kalamaras J, Hartigan-O'Connor DJ, Meneses JJ, Pedersen RA, et al. 1998. Mice lacking the homeodomain transcription factor Nkx2.2 have diabetes due to arrested differentiation of pancreatic beta cells. *Development* 125: 2213-21

Svensson P, Williams C, Lundeberg J, Ryden P, Bergqvist I, Edlund H. 2007. Gene array identification of Ipf1/Pdx1-/- regulated genes in pancreatic progenitor cells. *BMC Dev Biol* 7: 129

Tanaka M, Yamasaki N, Izumo S. 2000. Phenotypic characterization of the murine Nkx2.6 homeobox gene by gene targeting. *Mol Cell Biol* 20: 2874-9

Thomas PQ, Brown A, Beddington RS. 1998. Hex: a homeobox gene revealing peri-implantation asymmetry in the mouse embryo and an early transient marker of endothelial cell precursors. *Development* 125: 85-94

Thompson W, McCue LA, Lawrence CE. 2005. Using the Gibbs motif sampler to find conserved domains in DNA and protein sequences. *Curr Protoc Bioinformatics* Chapter 2: Unit 2 8

Thompson W, Palumbo MJ, Wasserman WW, Liu JS, Lawrence CE. 2004. Decoding human regulatory circuits. *Genome Res* 14: 1967-74

Thompson W, Rouchka EC, Lawrence CE. 2003. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res* 31: 3580-5

Thompson WA, Newberg LA, Conlan S, McCue LA, Lawrence CE. 2007. The Gibbs Centroid Sampler. *Nucleic Acids Res* 35: W232-7

Trinh DK, Zhang K, Hossain M, Brubaker PL, Drucker DJ. 2003. Pax-6 activates endogenous proglucagon gene expression in the rodent gastrointestinal epithelium. *Diabetes* 52: 425-33

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

van de Wetering M, Sancho E, Verweij C, de Lau W, Oving I, et al. 2002. The beta-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. *Cell* 111: 241-50

van den Brink GR. 2007. Hedgehog signaling in development and homeostasis of the gastrointestinal tract. *Physiol Rev* 87: 1343-75

van den Brink GR, Bleuming SA, Hardwick JC, Schepman BL, Offerhaus GJ, et al. 2004. Indian Hedgehog is an antagonist of Wnt signaling in colonic epithelial cell differentiation. *Nat Genet* 36: 277-82

van Es JH, van Gijn ME, Riccio O, van den Born M, Vooijs M, et al. 2005. Notch/gamma-secretase inhibition turns proliferative cells in intestinal crypts and adenomas into goblet cells. *Nature* 435: 959-63

Varnat F, Heggeler BB, Grisel P, Boucard N, Corthesy-Theulaz I, et al. 2006. PPARbeta/delta regulates paneth cell differentiation via controlling the hedgehog signaling pathway. *Gastroenterology* 131: 538-53

Wall ME, Dyck PA, Brettin TS. 2001. SVDMAN--singular value decomposition analysis of microarray data. *Bioinformatics* 17: 566-8

Wang J, Delabie J, Aasheim H, Smeland E, Myklebost O. 2002. Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. *BMC Bioinformatics* 3: 36

Wells JM, Melton DA. 1999. Vertebrate endoderm development. *Annu Rev Cell Dev Biol* 15: 393-410

Werner T. 2000. Computer-assisted analysis of transcription control regions. Matinspector and other programs. *Methods Mol Biol* 132: 337-49

Werner T. 2001a. Cluster analysis and promoter modelling as bioinformatics tools for the identification of target genes from expression array data. *Pharmacogenomics* 2: 25-36

Werner T. 2001b. The promoter connection. *Nat Genet* 29: 105-6

Werner T. 2002a. Finding and decrypting of promoters contributes to the elucidation of gene function. *In Silico Biol* 2: 249-55

Werner T. 2002b. Promoter analysis. *Ernst Schering Res Found Workshop*: 65-82

Werner T. 2007. Regulatory networks: linking microarray data to systems biology. *Mech Ageing Dev* 128: 168-72

Werner T. 2008. Bioinformatics applications for pathway analysis of microarray data. *Curr Opin Biotechnol* 19: 50-4

Werner T, Fessele S, Maier H, Nelson PJ. 2003. Computer modeling of promoter organization as a tool to study transcriptional coregulation. *FASEB J* 17: 1228-37

Werner T, Nelson PJ. 2006. Joining high-throughput technology with in silico modelling advances genome-wide screening towards targeted discovery. *Brief Funct Genomic Proteomic* 5: 32-6

Wu Z, Irizarry RA. 2004. Preprocessing of oligonucleotide array data. *Nat Biotechnol* 22: 656-8; author reply 8

Wu Z, Irizarry RA. 2005. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J Comput Biol* 12: 882-93

Xia X, Xie Z. 2001. AMADA: analysis of microarray data. *Bioinformatics* 17: 569-70

Yang Q, Bermingham NA, Finegold MJ, Zoghbi HY. 2001. Requirement of Math1 for secretory cell lineage commitment in the mouse intestine. *Science* 294: 2155-8

Yauch RL, Gould SE, Scales SJ, Tang T, Tian H, et al. 2008. A paracrine requirement for hedgehog signalling in cancer. *Nature* 455: 406-10

Zacchetti G, Duboule D, Zakany J. 2007. Hox gene function in vertebrate gut morphogenesis: the case of the caecum. *Development* 134: 3967-73

Zavros Y. 2008. The adventures of sonic hedgehog in development and repair. IV. Sonic hedgehog processing, secretion, and function in the stomach. *Am J Physiol Gastrointest Liver Physiol* 294: G1105-8

Zhang XM, Ramalho-Santos M, McMahon AP. 2001. Smoothened mutants reveal redundant roles for Shh and Ihh signaling including regulation of L/R symmetry by the mouse node. *Cell* 106: 781-92

Zhou XP, Woodford-Richens K, Lehtonen R, Kurose K, Aldred M, et al. 2001. Germline mutations in BMPR1A/ALK3 cause a subset of cases of juvenile polyposis syndrome and of Cowden and Bannayan-Riley-Ruvalcaba syndromes. *Am J Hum Genet* 69: 704-11

Zorn AM, Wells JM. 2007. Molecular basis of vertebrate endoderm development. *Int Rev Cytol* 259: 49-111

Zwijsen A, Verschueren K, Huylebroeck D. 2003. New intracellular components of bone morphogenetic protein/Smad signaling cascades. *FEBS Lett* 546: 133-9

CHAPTER II

DECONVOLUTING THE INTESTINE: MOLECULAR EVIDENCE FOR A MAJOR ROLE OF THE MESENCHYME IN THE MODULATION OF SIGNALING CROSSTALK

ABSTRACT

Reciprocal crosstalk between the endodermally-derived epithelium and the underlying mesenchyme is required for regional patterning and proper differentiation of the developing mammalian intestine. Though both epithelium and mesenchyme participate in patterning, the mesenchyme is thought to play a prominent role in the determination of the epithelial phenotype during development and in adult life. However, the molecular basis for this instructional dominance is unclear. In fact, surprisingly little is known about the cellular origins of many of the critical signaling molecules and the gene transcriptional events that they impact. Here, we profile genes that are expressed in the separate mesenchymal and epithelial compartments of the perinatal mouse intestine. The data indicate that the vast majority of soluble inhibitors and modulators of signaling pathways such as Hedgehog, Bmp, Wnt, Fgf and Igf are expressed predominantly or exclusively by the mesenchyme, accounting for its ability to dominate instructional crosstalk. We also catalog the most highly enriched transcription factors in both compartments. The results bolster previous evidence suggesting a major role for Hnf4α and Hnf4γ in the regulation of epithelial genes. Finally, we find that while epithelially

enriched genes tend to be highly tissue-restricted in their expression, mesenchymally enriched genes tend to be broadly expressed in multiple tissues. Thus, the unique tissue-specific signature that characterizes the intestinal epithelium is instructed and supported by a mesenchyme that itself expresses genes that are largely non-tissue specific.

Keywords:

Microarray, transcription factors, intestinal development, Bmp pathway

INTRODUCTION

The mammalian small intestine develops from a tube of endoderm wrapped by mesoderm. Throughout embryonic, fetal and adult life, the endoderm and mesoderm are mutually dependent upon one another for instructive signals that sculpt the eventual morphological form of each layer and control region-specific gene expression (Ratineau et al 2003). Elegant tissue recombination studies have underlined the extent to which this tissue crosstalk controls patterning. For example, in trans-species grafts of mouse and chick, the mesoderm holds the instructional information that controls whether the epithelium adopts villus projections as in the mouse or ridge-like structures characteristic of the chick (Simon-Assmann & Kedinger 1993). Likewise, dissociated ileal mesenchyme of the rat is able to instruct isolated rat colonic epithelium to express small intestinal enzymes (Duluc et al 1994), again reflecting the instructional power of the mesodermal layer. At specific timepoints during development, instructions also pass from endoderm to mesoderm, since ileal endoderm can induce colonic mesoderm to express homeobox genes characteristic of the ileal region (Duluc et al 1997). The accumulated data suggest that a mutually reinforcing program directs proper intestinal development, and predict that a predominant role is played by the mesenchyme in instructing and maintaining intestinal form and homeostasis (Kedinger et al 1998a). This prediction is further supported by the finding that different myofibroblast primary cell lines isolated from the gut induce different epithelial morphologies when co-cultured with an epithelial cell line such as CaCo2 (Duluc et al 1997).

Clearly, many complex signaling events regulate the development and regional patterning of the gut tube; signaling crosstalk also controls homeostasis as well as reaction to injury or inflammation. Microarray studies can shed light in a global way on subsets of genes that are changed in disease states or after alterations of specific signaling pathways in animal models. However, in many cases, microarray studies are performed on whole intestinal tissue, and thus components of both mesenchyme and epithelium are sampled together. It becomes difficult to decipher direct from indirect effects of any given signal in the absence of spatial information regarding the location of signaling molecules and transduction machinery, information which is not trivial to acquire, especially for multifactorial signals.

In this study, we have taken the first step in unraveling some of this complexity by examining gene expression profiles in separated epithelium and mesenchyme in the perinatal mouse intestine. At E18.5 (Figure 2.1A), intestinal villi are well-developed and most of the adult cell lineages are present (Paneth cells develop after birth). Villi are polarized, with the proliferative compartment restricted to the base of the villi (Figure 2.1B), but crypts have not yet formed. Clean separation of the epithelium from the mesenchyme is possible at this stage (Figure 2.1C, 2.1D).

This microarray analysis of isolated epithelial and mesenchymal fractions yields a molecular confirmation for a major role for the mesenchyme in the modulation of a wide variety of signaling pathways (e.g., Hedgehog, Wnt, Bmp, Fgf, Igf) since, with very few

exceptions, the soluble signal inhibitors and modulators of these pathways are expressed primarily in mesenchyme. We also identify all transcription factors that are expressed predominantly in one tissue compartment or the other and document substantial differences in the distribution of transcription factor sub-types (e.g., zinc finger, homeobox, etc.) in the two compartments. Finally an examination of the expression patterns of the most enriched epithelial and mesenchymal genes using electronically available EST tallies reveals that while many of the highly enriched epithelial genes are intestine-specific, nearly all genes highly enriched in the intestinal mesenchyme are expressed in numerous additional tissues. This suggests that a complex and multifactorial instructional signal from the mesenchyme is responsible for the generation and maintenance of a tissue-specific epithelial signature.

MATERIALS AND METHODS

Separation of epithelium and mesenchyme

All animals used in this study were C57BL/6; all experimental protocols performed on animals were done with previous approval from the University of Michigan Committee on the Use and Care of Animals (Protocol #7788). Tissue separation was done as previously described (Madison et al 2005). Briefly, E18.5 embryos from C57BL/6 females were removed from the yolk sac and amnion and placed into ice-cold PBS containing penicillin/streptomycin. The entire small intestine, from duodenum to cecum, was isolated, cut open longitudinally, placed into 1.0 ml of cold Cell Recovery Solution (BD Biosciences) in a 12-well plate and incubated for nine hours at 4°C. On ice, each

plate was agitated by hand for 30 to 45 minutes until all epithelium sloughed off, as determined by microscopic examination. The mesenchyme was gently removed from the more delicate epithelial fragments with sterile forceps and rinsed in a dish of sterile PBS to remove any remaining epithelial cells. Mesenchyme from three embryos was pooled into a 15 ml conical tube then snap frozen in liquid nitrogen; six such pooled samples were prepared. The remaining epithelial tissue was washed twice with 10 ml of cold PBS and collected by gentle centrifugation (100xg for 7 minutes at 4°C). Intestinal epithelium from three embryos was pooled into a 15 ml conical tube then snap frozen in liquid nitrogen; six pooled samples were prepared. Each pool of tissue was subsequently homogenized in 1.0 mls TRIzol (Invitrogen), on ice, and RNA prepared per the TRIzol protocol. Total RNA was further purified using the RNeasy Mini Kit (Qiagen), and quality assessed by electrophoresis of 2.0 μg RNA on a 1% agarose gel. To assess epithelial vs. mesenchymal purity of RNA, we performed RT-PCR for both epithelial-specific mRNAs (Vil1 and Ihh) and mesenchymal-specific mRNAs (Madcam1, and Actg2). Three epithelial and three mesenchymal RNA fractions exhibiting minimal contamination with mesenchymal and epithelial mRNAs, respectively, were selected from among all of the samples for Affymetrix microarray analysis. According to the manufacturer's instructions, labeled cRNA probes were synthesized from total purified RNA and hybridized to the Affymetrix GeneChip® Mouse Genome 430 2.0. The University of Michigan Diabetes Center Microarray Core Facility performed cRNA synthesis, array hybridization, and array scanning, all according to standard Affymetrix protocols.

Microarray Analysis

Microarray was performed using Affymetrix MOE 430.2 arrays, containing approximately 43,000 probes to study over 39,000 characterized genes or ESTs. Six microarray chips were used (three epithelium and three mesenchyme) for the analysis. After the hybridizations, microarrays were washed and scanned to generate the image files (.DAT files), which were then processed using MAS 5.0 software to produce .CEL files. The .CEL files were analyzed using RMA (Robust Multi-Array Average) which subtracted background, normalized expression data and calculated gene expression values(Irizarry et al 2003b). The logarithm base 2 (Log_2) of the average expression value for each probe pair on the three epithelial chips was subtracted from the Log_2 of the average expression value on the three mesenchymal chips. The difference was converted to a numerical Fold Difference ($FD = 2^{Log2(mes)-Log2(epi)}$) or Enrichment Score (ES, absolute value of the Fold Difference). Two-tailed T tests were carried out to identify differentially expressed genes according to p-value ≤ 0.05 and fold difference ≥ 2.0. Where multiple probe sets were available for a given gene, only the probe set with the highest enrichment score is listed in the associated data tables. All data are available at the NCBI GEO Database (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6383) under the series number GSE6383.

Search for Hnf4 binding sites

Promoter sequences (500 bp upstream of transcriptional start site) were downloaded from Ensembl (http://www.ensembl.org/Mus_musculus/) for each of the 100 most enriched genes in the epithelium and each of the 100 most enriched genes in the mesenchyme. In addition, three groups of 100 genes with enrichment scores of 1.1 - 0.9 were compiled and promoter sequences (500 bp) were downloaded. These three groups differed in their average expression level: CntlH included 100 genes with average expression values greater than 10.0 in epithelium and mesenchyme; CntlM included 100 genes with average expression levels between 8.0 and 10.0 in both compartments; CntlL consisted of 100 genes with average expression values less than 8.0 in epithelium and mesenchyme. The MatInspector tool of the Genomatix software suite (http://www.genomatix.de/) was used to search for Hnf4 binding sites in these five groups of genes (Cartharius et al 2005, Quandt et al 1995). The default stringency settings in MatInspector were used in the search. MatInspector uses two consensus Hnf4 sequences based on functional studies in the literature (Fraser et al 1998): $R_{31}G_{87}G_{59}N_{16}C_{100}A_{100}A_{100}A_{100}G_{100}K_{50}T_{37}C_{67}A_{61}$ (Hnf4.01) and $S_{57}R_{57}G_{100}G_{100}T_{65}C_{100}M_{57}A_{100}A_{100}A_{100}G_{100}G_{65}T_{100}C_{100}$ (Hnf4.02). The subscripted numbers indicate the conservation value assigned to each nucleotide. Bases with a conservation value >60 are considered to have a high information content.

RT-PCR verification

RT-PCR was used for quality control on the original separated tissue samples, and to verify results of the microarray. All RT-PCR was done using cDNA produced from

isolated epithelial or mesenchymal RNA. Conditions for amplification (magnesium, DMSO addition) were individually optimized for each probe set. Unless otherwise indicated, 30 cycles of amplification were used. Primers and conditions used for all PCR studies are listed in Table 2.1.

In situ hybridization

In situ probes were synthesized from cDNA fragments obtained by rtPCR and cloned into PCR4 TOPO vectors (Invitrogen). The following templates were used: Bmp2 (NM_007553 nucleotides 220-1250), Bmp4 (NM_007554 nucleotides 1-1115), Bmp5 (NM_007555 nucleotides 75-950), Bmp7 (NM_007557 nucleotides 770-1780), Bmpr2 (AF003942 nucleotides 3410-2595), Bmpr1a (NM_009758 nucleotides 390-1450) and Acvr1 (NM_ 007349 nucleotides 135-1165). To ensure the specificity of the probes, both sense and antisense probes were generated from each template and tested in parallel.

Small intestines from E17.5 C57Bl/6 embryos were removed and fixed overnight in 4% paraformaldehyde. The tissue was then dehydrated, processed and embedded in paraffin, sectioned at 5 μm and baked for 2 h at 56°C. For in situ hybridization, sections were dewaxed, rehydrated, digested in a 20ug/ml proteinase K solution for 10 minutes at 37°C, postfixed and treated in acetic anhydride solution. After prehybridization in 4 x SSC, 50% formamide at 37°C for 1h, the sections were hybridized overnight at 60°C with various probes in 50% formamide, 10% dextran sulfate, 1mg/ml yeast tRNA, 1x

Denhardt's, 0.2M NaCl, 10mM Tris-HCl pH 7.5, 10mM phosphate buffer pH 6.8 and 5 mM EDTA pH 8.0. After hybridization, the slides were washed for 3 x 30 minutes in 1X SSC, 50% formamide, 0.1% Tween 20 at 65°C, and then for 2 x 30 minutes in TBST (50 mM Tris-HCl pH 7.5, 150 mM NaC, 0.1 % Tween 20) at room temperature. Sections were then blocked for 1 h in TBST containing 20% heat inactivated FCS and 2% blocking reagent (Roche) and incubated in blocking solution with alkaline phospatase-conjugated anti-digoxigenin antibody (Roche), 1:1000 dilution at 4°C over night. After several washes in TBST the slides were equilibrated in staining buffer, 100 mM NaCl, 50mM MgCl2, 100mM Tris-HCl pH 9.5, 0.1% Tween 20. The color reaction was performed in 10% polyvinyl alcohol, 100 mM NaCl, 5mM MgCl2, 100mM Tris-HCl pH 9.5, 0.1% Tween 20, 100mg/ml NBT and 50mg/ml BCIP for 10-20 hours. The color reaction was stopped in PBS and the slides mounted with 70% glycerol/PBS.

RESULTS

Validation of clean tissue separation

The entire small intestine (duodenum to cecum) was removed from E18.5 fetuses and separated into epithelial and mesenchymal fractions as shown in Figure 2.1C and 2.1D (Madison et al 2005). To assess the efficacy of the separation, cDNA was prepared from each individual fraction and analyzed by RT-PCR using primers for genes known to be expressed exclusively in either the epithelial (Villin1, Ihh) or mesenchymal (Actg1, MAdCAM) compartment. Only fractions that appeared to be free of contamination from the other compartment were used for microarray analysis (Figure 2.1E).

Six microarray chips (three mesenchymal mRNA and three epithelial mRNA) were hybridized using independently isolated samples. The difference in relative expression of each probeset was calculated as described in Materials and Methods and expressed as a numerical Fold Difference (FD) or Enrichment Score (ES, equal to the absolute value of the Fold Difference). In Figure 2.2A, all data are plotted according to [$-\text{Log}_{10}$(P value), y axis] vs. [Log_2(Fold Difference), x axis] in a volcano plot. The boundaries of Fold Difference ≥ 2.0 and p-value ≤ 0.05 are indicated in the figure; these boundaries demarcate genes that are enriched in the epithelium (EPI: 1812 known genes; 111 unknown ESTs or cDNAs) and in the mesenchyme (MES: 4245 known genes; 417 unknown ESTs or cDNAs). Top 100 epithelially and mesenchymally enriched genes, along with their average relative expression levels and enrichment scores (ES) are provided in Tables 2.2 and 2.3, respectively. Complete lists of epithelially and mesenchymally enriched genes are provided in Supplementary Tables 2 and 3, respectively on http://physiolgenomics.physiology.org/cgi/content/full/00269.2006/DC1; or by searching PMID: 17299133.

The average relative expression value in epithelium (y axis) vs. mesenchyme (x axis) for each of the queried probes in the array as determined from the RMA analysis is graphically depicted in Figure 2.2B. The data are well spread over approximately 11 relative expression units (2^5-2^{15}). We examined the epithelial and mesenchymal expression levels for several housekeeping genes that are known to be ubiquitously

expressed: Hprt, Gapdh, Ppig (cyclophilin G) and Actb. In each case, the enrichment score is close to 1.0 (Table 2.4). Thus, these genes will fall on the diagonal in Figure 2.2B.

Additionally, we queried the data for several genes known to be compartment specific, including structural genes, transcription factors and secreted factors: Vil1, Fabp1, Cdx1, Cdx2 and Ihh for the epithelium; Vim, Actg2, Foxf1, Gli1, and Bmp4 for the mesenchyme. The results (Table 2.4) show robust compartment-specific enrichment scores for the genes that are expressed at high levels (e.g., Villin1 in the epithelium; Vimentin in the mesenchyme). However, for genes expressed at lower levels or in few cells (e.g., Ihh in the epithelium), the enrichment scores are lower (Table 2.4). This is a consequence of the low positive expression value in one compartment coupled with a relatively constant background signal in the other compartment of 4-5 relative expression units. Thus, low enrichment values (e.g., for Ihh, ES = 2.8) can still be indicative of compartment-specific expression. Indeed, experimental data obtained using in situ hybridization, immunohistochemistry and PCR have confirmed that both Ihh and Shh (ES = 1.9) are in fact exclusively epithelial (Madison et al 2005, Ramalho-Santos et al 2000). Thus, the array results present an accurate prediction of epithelial vs. mesenchymal compartmentalization. Nevertheless, some genes that are expressed at low levels or in few cells will be missed at a cut-off value of ES = 2.0, even if they are (like Shh), truly compartment specific.

Functional attributes of genes expressed in mesenchyme and epithelium

As a second check on the reliability of the array data, we tallied the Gene Ontology (GO) terms linked to genes with FC > 5.0 and $p \geq 0.05$. GO terms provide a "dictionary" of labels that describe three aspects of each gene: the Biological Process term reflects the cellular processes to which each gene is linked (e.g., signaling, proteolysis); the Cellular Component term describes the cellular organelle with which the protein is associated (e.g., cytosol, nucleus) and the Molecular Function term denotes the function performed by the transcribed protein (e.g., transcription factor, kinase).

As shown in Figure 2.3, the Biological Process terms differ strikingly between the two compartments in a way that reflects their known functions. Among epithelially enriched genes, nearly 50% have linked Biological Process terms that reflect a role in nutrient absorption and processing: metabolism (21%), transport (20%), and proteolysis (7%). In contrast, signal transduction (12%) and transcription (12%) are the most frequently linked Biological Process GO terms for the mesenchymally enriched genes. In concert with this, the most prominent Cellular Component GO term for the epithelium (membrane, 33%) reflects its role in absorption and transport, while the predominant Cellular Component term for the mesenchyme ("extracellular", 33%) is in keeping with the involvement of the mesenchyme in extracellular matrix production and instructional signaling.

A tissue-specific signature held by the epithelium

We next examined the tissue-specificity of the 100 most enriched genes in the epithelium and mesenchyme (see Tables 2.2 and 2.3). These genes tend to be among the more highly expressed genes in their respective compartments. The tissue specificity of each gene was estimated by counting the number of different tissues from which ESTs for that gene have been isolated according to the EST ProfileViewer (Unigene, NCBI). This analysis shows that the epithelially-enriched genes are generally expressed in few tissues outside of the intestine while mesenchymally enriched genes are broadly expressed in multiple tissues (Figure 2.4). A total of 47 of the 100 most epithelially enriched genes are expressed in 10 or fewer tissues, while only five of the most mesenchymally enriched genes (Phox2, Adamdec1, Colec10, Cnn1, Sh3bgr) are this restricted.

Compartmentalized transcription factors in mouse intestine

Transcription factors play critical roles in controlling signaling cascades. A recent study revealed that over 1000 such factors are expressed in various regions of the gut tube during development(Choi et al 2006). From our epithelially and mesenchymally enriched gene sets (For top 100, see table 2.2 and table 2.3.Complete list in Supplementary Tables 2 and 3 on http://physiolgenomics.physiology.org/cgi/content/full/00269.2006/DC1.), we pinpointed all transcription factors that are enriched by at least 2.0-fold ($p \geq 0.05$) in each compartment. These genes were initially selected if they were tagged with the Biological Process term “transcription”, the Cellular Component “nucleus”, and/or the

Molecular Function term "DNA binding". Genes were then searched individually to either confirm experimental evidence for function related to Pol II transcriptional activity or to identify specific domains (e.g., Kruppel-type zinc finger domains) that are highly suggestive of transcriptional function. A total of 76 real and putative transcription factors were found to be enriched in the epithelium (Table 2.5) and 373 were identified as enriched in mesenchyme (Table 2.6). These 449 factors that are differentially expressed between epithelium and mesenchyme represent about 40% of the total number of transcription factors expressed in the developing gut, according to a recent study by Choi et al.(Choi et al 2006). The Choi analysis included both stomach and small intestine and the analysis was done at 4 time points during development (E11, E13, E15 and E17). Since our study only examines one time point (E18.5) and one gut region (small intestine), it is striking that so many factors are compartment-enriched.

The list of highly enriched epithelial transcription factors includes many with known roles in gut development and/or disease (e.g., Hnf4a, Hnf4g, Klf5, Ehf, Cdx1, Cdx2, Gata4, Gata5, Gata6, etc.). Nevertheless, among the 10 most highly enriched epithelial factors are several without previously described roles in the intestine. Tcfec, the third most epithelially enriched factor (ES = 18), is a bHLH-leucine zipper family member closely related to Mitf, Tfe3 and Tfeb, members of the micropthalmia class of transcription factors that play critical roles in eye and melanocyte development (Rehli et al 1999). Creb3l3 (Creb-H; ES = 11) is a bZip family member that is abundantly expressed in liver, where it appears to act as a suppressor of proliferation (Chin et al 2005). In accord with this, it is down-regulated in hepatomas (Chin et al 2005).

Similarly, Ehf (ESE-3; ES = 11), an Ets family member, is also modulated in cancer and has a postulated role in the differentiation of glandular epithelia (Kas et al 2000). The robust expression of these factors in the intestinal epithelium suggests that they would be interesting targets for further investigation.

Figure 2.5A provides RT-PCR data that verify transcription factor assignments made by the microarray for the 16 most enriched factors in the epithelium. These factors were between 66 and 6 fold enriched in epithelium, according to the array data, and the PCR results are in accord with these data. A few transcription factors with lower enrichment scores were also examined by RT-PCR: longevity assurance homolog 6 (Lass6, ES = 3.3); RAR-related orphan receptor gamma (Rorc, ES = 3.4); E2F transcription factor 5 (E2F5, ES = 3.0); and hairy and enhancer of split 6 (Hes6, ES = 2.1). In each case, PCR evidence of epithelial predominance was also obtained, indicating that even when the enrichment score is low, the array results accurately reflect compartmental enrichment.

Nearly 400 transcription factors were found to be at least two-fold enriched in the mesenchyme (Table 2.6). This is almost five times the number of transcription factors found to be enriched in the epithelium. Perusal of the mesenchymal list reveals factors known to be expressed in hematopoietic cells (Fli1), enteric neurons (Phox2), smooth muscle cells (myocardin) or myofibroblasts (Foxf1). Thus, the heterogeneous cell composition of the mesenchymal compartment may account for the large number of transcription factors that are enriched there.

Of the 10 most highly enriched factors in this compartment, three (Ets1, Elk3 and Fli1) are members of the Ets transcription factor family. These factors have overlapping gene expression patterns in fibroblasts, endothelial cells and hematopoietic cells. The active expression of all three of these Ets factors is in accord with an important mesenchymal participation in matrix remodeling and cell migration in the E18.5 intestine (Buchwalter et al 2005, Hsu et al 2004, Nakerakanti et al 2006).

Enrichment of transcription factor sub-types

Gray et al. recently mined the mouse genome to identify all transcription factors and categorized each factor according to its type of DNA binding domain (Gray et al 2004). We compared the frequency of each type of DNA binding factor in our collection of epithelial or mesenchymal transcription factors to the genomic frequency of each type as compiled in the Gray study. The classifications of each factor identified in the epithelium and mesenchyme are included in Tables 2.5 and 2.6 and the compiled results of the distribution analysis are presented in Figure 2.5B and 2.5C. The zinc finger (ZnF) class of transcription factors is the predominant class in the mouse genome (Gray et al 2004). Interestingly, the relative frequency of the various types of mesenchymally enriched transcription factors closely mirrors that of the entire mouse genome (Figure 2.5B). However, the epithelially enriched factors exhibit a very different profile, with an under-representation of ZnF and HMG (high mobility group) factors and an overrepresentation of HLH (helix loop helix), bZip and NHR (nuclear hormone receptor) classes (Figure 2.5C). The prominence of NHR factors in the epithelium is particularly

striking. These proteins make up only 2% of the mesenchymally enriched transcription factors, but represent 17% of the epithelially enriched factors (Figure 2.5B).

Hnf4 binding sites in epithelial genes

Hnf4a and Hnf4g are the most highly enriched transcription factors in the epithelium. In fact, Table 2.2 reveals that both Hnf4a and Hnf4g rank very high (#7 and #22, respectively), in the total list of epithelially enriched molecules. Hnf4a has been shown to bind to a large number of promoters in the liver (Odom et al 2004) and functional studies indicate that it is essential for epithelial differentiation in the liver(Parviz et al 2003) and the colon(Garrison et al 2006). If Hnf4 proteins are also key regulators of transcription in the small intestinal epithelium, then a preponderance of epithelially enriched genes should exhibit cis binding sites for Hnf4. We downloaded 500 bp of sequence upstream of Transcription Start Site (TSS) from the 100 most enriched epithelial genes and the 100 most enriched mesenchymal genes from Ensembl. In some cases, more than one promoter (more than one transcriptional start site or TSS) was identified for a single gene. Thus, a total of 114 and 123 sequences were downloaded and searched for the top 100 epithelial and 100 mesenchymal genes, respectively. MatInspector from the Genomatix software suite (www.genomatix.de) was used to search for potential Hnf4 binding sites in these sequences. As outlined in Materials and Methods, this program searches for two variants of the consensus Hnf4 binding site, both verified by functional binding site selection and affinity studies(Fraser et al 1998). When a single site was identified by both motifs, it was counted only once. To establish a baseline for

the frequency of finding these cis elements in a similar size set of random genes, we also examined three groups of 120 sequences from among the genes that showed no enrichment between epithelium and mesenchyme (ES = 1.0). The three groups included genes expressed at high (H), medium (M) and low (L) relative expression values, as described in Materials and Methods.

Using the default match criteria of MatInspector, 61 sequences from 114 epithelially enriched genes were found to contain putative Hnf4 sites (listed in Table 2.7). In contrast, Hnf4 sites were found in only 29 sequences among the 123 sequences from mesenchymally enriched genes. Among the four groups of 120 sequences with enrichment scores of 1.0, the number of Hnf4 binding sites ranged from 23-37, similar to the number seen in mesenchymally enriched genes and approximately half of the number seen in epithelially enriched genes (Figure 2.6A). In addition, the actual number of Hnf4 binding sites found was 42 in mesenchymally enriched genes and 85 in the epithelial set. Control groups ranged from 29-56 binding sites, similar to the number seen in the mesenchymally enriched set and half of the number seen in the epithelial set. Increasing the stringency of the Hnf4 match does not alter these findings, since 20 of the binding sites in epithelial genes have a better than 90% match to the consensus sequences (Table 2.7, see Matrix Similarity). However, only two sites in the mesenchyme match at this level (data not shown).

We next examined the location of consensus Hnf4 binding sites within the 500 bp promoters of mesenchymally and epithelially enriched genes containing such sites. Interestingly, Hnf4 binding sites appeared to be strikingly scarce in the first 100 bp of promoter sequence of all three control groups (solid symbols in Figure 2.6B) and randomly distributed over the 200-500 bp range of these sequences. Importantly, the mesenchymally enriched genes (open triangles in Figure 2.6B) displayed a pattern very similar to those of the three control groups. In contrast, consensus Hnf4 binding sites were highly concentrated in the first 300 bp of the promoters of epithelial genes (Figure 2.6B, open diamonds). The number of binding sites was particularly high in the region between 0 and -100 bp from the transcriptional start site, the same region in which such sites appear to be scarce in mesenchymal and control genes. Taken together, these results predict that Hnf4 regulates a large number of intestinal epithelial genes, as in liver and colon. These data are in accord with those of Stegmann et al. who found Hnf4 sites in a large number of intestinal genes that are highly up-regulated during development of the intestine from the early fetal (E13) to adult stage and/or are up-regulated in adult villus tips compared to crypts (Stegmann et al 2006).

The mesenchyme as a modulator of signal transduction

To learn more about the direction of signaling crosstalk in the intestine, we next examined the enrichment values for molecules that participate in several signaling pathways known to be important in intestinal development and homeostasis, including:

Notch, Hedgehog (Hh), Wnt, Igf, Fgf and Bmp. Genes enriched more than two fold in either compartment are listed in Tables 2.8 and 2.9.

Notch signaling in the epithelium is essential for the control of epithelial lineage allocation (Milano et al 2004, Stanger et al 2005, van Es et al 2005) and recent work has also implicated Notch signaling in the expansion of the crypt progenitor pool(Fre et al 2005). The array data collected here are consistent with low levels of Notch expression in the epithelium (average expression values of 7.3 to 9.5), but it is interesting that all four of the Notch genes are actually more enriched in the mesenchyme (Table 2.8): Notch1 (ES = 3.1), Notch2 (ES = 6.2), Notch3 (ES = 4.8) and Notch4 (ES = 6.4). Thus, it is likely that important mesenchymal roles for this pathway also exist.

Current experimental data indicate that Hh signals originate in the epithelium and that the Hh signal transduction machinery is located exclusively in the mesenchyme (Madison et al 2005). The microarray data are consistent with this (Table 2.8). Though enrichment values for Ihh and Shh in the epithelium are low, most likely because of their expression in few cells, significant mesenchymal enrichment is seen for the receptors Ptc1 (ES = 27.7) and Ptc2 (ES = 3.4), co-receptor Smo (ES = 4.9) and transcription factors Gli1 (ES = 6.3), Gli2 (ES = 3.4) and Gli3 (ES = 9.3). A single pan-inhibitor of the Hh signal has been described: Hh interacting protein (Hhip) (Chuang & McMahon 1999). Like Ptc, this membrane-bound inhibitor binds Hh with high affinity and is also a direct target of Hh signaling. Hhip expression was earlier shown to be restricted to the intestinal

mesenchyme (Madison et al 2005); in agreement with this, the array data suggests a 25 fold enrichment of Hhip in the mesechmymal compartment.

Information concerning the spatial distribution of several Wnt signaling molecules in the developing and adult intestine has been previously published (Gregorieff et al 2005, Lickert et al 2001, McBride et al 2003, Theodosiou & Tabin 2003). The microarray data for E18.5 fetal intestine largely corroborates those studies, with a few interesting differences (see Table 2.8). First, the non-canonical receptor, Fzd6, appears to be expressed in both compartments of the late fetal intestine but is somewhat more enriched in the mesenchyme (ES = 3.3). In contrast, in adult intestine, expression of this factor is clearly predominant in the epithelium (Gregorieff et al 2005). Fzd1, Fzd2 and Fzd7 are also strongly mesenchymally enriched in our E18.5 samples (ES = 8.0, 15.6 and 10.6, respectively). This correlates well with the adult findings for Fzd1 and Fzd2, which are both expressed in smooth muscle, but not for Fzd7, which is epithelial in the adult (Gregorieff et al 2005). Finally, the inhibitors Dkk-2 and Sfrp-2 are highly mesenchymally enriched in the E18.5 intestine (ES = 10 and 11.7, respectively), but were not detected in adult intestine (Gregorieff et al 2005). These differences suggest a temporally dynamic patterning of Wnt pathway molecule expression.

It is striking that Wnt5a (ES = 15.3) and Fzd2 (ES = 15.6) exhibit the highest enrichment values for any Wnt family member and Frizzled receptor, respectively and both are enriched in mesenchyme. Both of these proteins can function in the non-canonical Wnt

signaling pathway, as can Fxd6, which is enriched 3.3 fold in the mesenchyme. These findings suggest that beta-catenin-independent signaling may play an important role in this compartment.

As seen with the Hh signaling system, the array data indicate that the predominant site for manufacture of inhibitors of the Wnt pathway is the mesenchyme. Sfrp1, Sfrp2, Dkk2 and Dkk3 are all greatly enriched in this compartment (ES = 78.5, 11.7, 10 and 8, respectively). The relative expression values suggest that Sfrp4 and Sfrp5 are expressed at low to moderate levels in both epithelium and mesenchyme (Supplementary Tables 2 and 3 on http://physiolgenomics.physiology.org/cgi/content/full/00269.2006/DC1), but strikingly, none of the Wnt inhibitors are preferentially enriched in the epithelial compartment.

In the Igf signaling system, several Igf binding proteins (Igfbp) act to inhibit or potentiate Igf activity, often in a context-specific manner(Cohen 2006). Like the Wnt and Hh inhibitors, Igfbps are prominently expressed in the mesenchyme and in fact are among the most highly expressed and highly mesenchymally enriched of any of the signaling molecules analyzed (Table 2.8). Enrichment values of over 20 fold are seen for Igfbp2, Igfbp3, Igfbp4, Igfbp5 and Igfbp7. Messenger RNA for Igfbp3, the binding protein to which the majority of circulating Igf is bound(Guler et al 1987), is over 100 fold enriched in mesenchyme and has a very low relative expression value (5.7) in epithelium. This protein has been shown to possess both Igf-dependent and Igf-independent activities and

has pro-apopototic activity in the absence of Igf (Hong et al 2002). Igf1 as well as the receptor Igf1r (which binds both Igf1 and Igf2) are also enriched in mesenchyme.

Fgf signaling has been shown to play a role in anterior/posterior patterning of the early gut tube (Dessimoz et al 2006) but its roles in later gut development are unclear. By far, the most robustly expressed Fgf family member in the E18.5 day intestine is Fgf13, which is enriched 47.6 fold in mesenchyme (Table 2.8). Many other Fgfs are expressed at low levels in both epithelium and mesenchyme (relative expression values of 7 or greater), including Fgf1, Fgf3-15, Fgf17, Fgf18, Fgf21 and Fgf22 (data not shown). However, only Fgf5, 7 and 15 are enriched in a particular compartment, with Fgf5 and Fgf15 enriched in epithelium (ES = 2.3 and 2.9, respectively) and Fgf7 enriched 2.2 fold in mesenchyme (Table 2.8).

Two findings suggest that Fgf signaling may be particularly important in mesenchyme. First, of the four Fgf receptors, Fgfr1 and Fgfr2 are highly expressed in mesenchyme and are enriched by 35.5 and 3.7 fold, respectively, in that compartment. Fgfr3 and Fgfr4 are expressed at much lower levels, likely in both epithelium and mesenchyme (data not shown). Second, several members of the Sprouty (Spry) and Sprouty-related (Spred) families are expressed predominantly in the mesenchyme. These proteins are intracellular inhibitors of Fgf signaling and Spry2 is a direct target of Fgf signals(Chambers & Mason 2000). As shown in Table 2.8, Spry1, 2 and 4 are enriched 5, 7 and 35 fold in mesenchyme, respectively; Spred1 and Spred2 are also mesenchymally enriched (ES =

13.6 and 3, respectively). Interestingly, Fgfbp1, a soluble enhancer of Fgf signals, is one of the few signaling modulators that we found to be enriched in epithelium (ES = 6.2). Fgfbp1 binds directly to several Fgfs and potentiates their activity(Beer et al 2005).

The Bmp pathway: modulation by numerous mesenchymal factors.

Recent investigations have suggested that Bmp signaling to the epithelium is important in patterning the crypt/villus axis, and in pathological states such as juvenile polyposis syndrome or JPS (Haramis et al 2004, He et al 2004, Howe et al 2001). However, the expression of Bmp pathway molecules other than Bmp4 has not been systematically examined in the intestine; thus we mined the array data for compartmentalized Bmp pathway molecules (Table 2.9).

Bmp4 is highly expressed and enriched (ES = 23.8) in the mesenchyme, in agreement with previous reports (Haramis et al 2004, He et al 2004, Madison et al 2005). But this is not the only Bmp family member that is enriched in this tissue layer: Bmp5 is 25 fold enriched, Bmp6 is 3.3 fold enriched and Bmp2 is 2.2 fold enriched in mesenchyme. Bmp7 is the only family member showing epithelial enrichment (ES = 4.7). PCR validation of these results is provided in Figure 2.7A. In situ hybridization studies further confirmed these expression patterns (Figure 2.7D). Particularly noteworthy here is the pattern of Bmp5, which, like Bmp4, is highly expressed in stromal cells adjacent to the epithelium.

The array data suggest that the receptors and signal transducers (Smads) of the Bmp pathway are expressed in both epithelial and mesenchymal compartments (Table 2.9 and see the entire original raw microarray data on NCBI GEO database at (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6383) with the series record number GSE6383). Bmpr2, the most robustly expressed of all of the receptors, is enriched 4 fold in mesenchyme, though it is also expressed in the epithelium (Table 2.9; Figure 2.7D). Two of the three type I receptors, Bmpr1a (Alk3) and Acvr1 (Alk2), are mesenchymally enriched (2.6 fold and 4.5 fold, respectively), while the third, Bmpr1b (Alk6) is expressed at much lower levels in both mesenchyme and epithelium and not enriched in one particular compartment (data not shown). In situ hybridization confirms mesenchymal enrichment as well as expression in the epithelium at the base of the villi for Bmpr2 and Acvr1 (Figure 2.7D). Bmpr1a is expressed primarily in the mesenchyme in the proximal duodenum (data not shown) but is found in both compartments in the more distal small intestine (Figure 2.7D). Of the Smads, Smad1, Smad4, Smad5, Smad6 and Smad7 are expressed and all except Smad4 are slightly enriched in the mesenchyme (Table 2.9; Figure 2.7). Smad4 exhibits similar, high relative expression values of 10-11 in both epithelium and mesenchyme. These data indicate that Bmp signals most likely impact both epithelial and mesenchymal compartments, though studies to date have primarily addressed epithelial signal transduction (He et al 2004).

The Bmp pathway is particularly rich in molecular inhibitors and modulators and these can act as on/off switches for Bmp function(Yanagita 2005). Strikingly, and consistent with findings for the Hh, Wnt, Fgf and Igf families, the expressed soluble inhibitors and modifiers of the Bmp signaling pathway are located primarily or exclusively in the mesenchyme, including: three follistatin family members (follistatin, follistatin 5 and follistatin-like 1), chordin-like 1, twisted gastrulation, Tolloid (Bmp1), Tolloid-like1, Bmper (Cv2), Gremlin1 and Crim1 (Figure 2.7, Table 2.9). This finding places the major control of Bmp signaling activity squarely in the domain of the mesenchymal compartment.

DISCUSSION

In this study, we established an intestinal tissue catalog wherein hundreds of intestinal genes are categorized according to their likely location in the epithelium and/or mesenchyme. Because the technique of tissue separation is very effective and little contamination is detectable between compartments (Figure 2.1 C-E), compartmentalization is clear and the data are verifiable by PCR. Within this dataset, all transcription factors that are enriched in epithelium or mesenchyme were identified. The expression of genes involved in several signaling pathways that mediate inter-tissue communication were also cataloged, providing a clearer picture of the contributions of each compartment in tissue crosstalk. These data will provide a useful guide that will contribute to the future dissection of developmental and pathological processes in the intestine.

Since heterotypic grafting experiments have shown that the mesenchyme holds a well-demonstrated power to instruct the regional specificity of the epithelium, both in the embryo and the adult (Kedinger et al 1998a), we were particularly interested to investigate whether the transcriptome of the epithelium and/or mesenchyme could suggest a molecular basis for this. We found that the epithelial transcriptome is highly tissue-restricted, as expected given its specialized functional roles. However, the underlying mesenchyme exhibits a rather ubiquitous-looking gene expression profile. Similarly, at the transcription factor level, the profile of transcription factor classes expressed by the epithelium differs markedly from the profile of the genome as a whole, while the mesenchymal transcription factor profile closely resembles that of the genome. It is possible that instructional signals from the mesenchyme are powered by the few more tissue-specific genes present in that tissue (e.g., genes such as Foxf1 and Nkx2.3). In support of this, the Nkx2.3 null mouse as well as the $Foxf1^{+/-}/Foxf2^{+/-}$ double heterozygote exhibit defects in mesenchymal as well as epithelial patterning (Ormestad et al 2006, Pabst et al 1999, Wang et al 2000). Alternatively or additionally, the instructional ability of the mesenchyme might rely on the use of a large arsenal of soluble signaling proteins in a combinatorial manner; indeed we show here that the mesenchymal transcriptome contains such an arsenal. Across several signaling pathways, (Notch, Hh, Wnt, Igf, Fgf and Bmp), a mesenchymal predominance is notable for several of the signaling molecules themselves (Table 2.8). But most striking for all of these pathways is the remarkable number of soluble inhibitors and modulators of signaling that appear to be expressed predominantly or exclusively in the mesenchyme. Thus, the

mesenchyme has an enormous potential to activate or suppress these important instructional pathways.

We paid particular attention to the Bmp pathway in this study since this pathway is important in both development and disease, yet few of the multiple molecular participants in this pathway have been carefully studied. The array data revealed that the mesenchyme has the ability to control multiple aspects of Bmp-mediated intestinal patterning and homeostasis. Among the Bmp ligands, both Bmp4 and Bmp5 are highly enriched in the mesenchymal compartment. Bmp5 has been less widely studied than Bmp4, but a mouse mutation has been identified that has an intestinal phenotype (*short ear, se*). *Se* mice exhibit defective skeletal structures and show changes in the morphology of several soft tissues, including the intestine, where intestinal looping is altered(Green 1968). Among Bmp inhibitors and modulators, robust enrichment of a remarkably large variety of these molecules is observed in the mesenchyme (Table 2.9, Figure 2.7). Given the fact that juvenile polyposis syndrome (JPS) (Howe et al 2004), a relatively rare autosomal dominant syndrome involving hamartomatous polyps that predisposes to gastrointestinal cancer, is known to involve alterations in Bmp pathway signaling, it will be important to test whether mutations in Bmp5 or any of these multiple Bmp pathway inhibitors can also give rise to this syndrome.

The distribution of Bmp receptors and Smads in epithelium and mesenchyme suggests that both compartments can receive and transduce Bmp signals. Particularly interesting is

the distribution of Smad4 and Bmpr1a, both of which are transcribed in epithelium as well as mesenchyme; Bmpr1a is actually enriched in the mesenchyme, particularly in the proximal small intestine, though Smad4 is not. Inactivating mutations in these two genes account for 40% of the cases of JPS (Howe et al 2004). Interestingly, there is still debate as to whether the epithelium or the mesenchyme is responsible for initiation of the JPS pathology. Early studies of patients with SMAD4 mutations revealed clonal genomic deletions only in the mesenchyme(Jacoby et al 1997). Thus, the epithelial malignancy was postulated to be due to landscaping by the mesenchyme. Later work identified loss of both SMAD4 alleles in the epithelium, a finding more consistent with a tumor suppressor mechanism(Woodford-Richens et al 2000). But interestingly, biallelic SMAD4 loss was found in some stromal and pericryptal fibroblasts of the polyp, suggesting a possible clonal origin for the mutant epithelium as well as part of the stroma of the polyp. The authors speculated that epithelial cells that lose the ability to receive or process Bmp signals might give rise to mesenchymal cells. In this light, it is interesting that epithelial/mesenchymal transition in the kidney is regulated by the balance of Bmp (promoting epithelial) and Tgf (promoting mesenchymal) signals(Wahab & Mason 2006). Though functional corroboration is needed, the data presented here are consistent with the possibility that the mesenchyme controls this balance and thus may be the compartment that is primarily responsible for the emergence of JPS pathology.

In summary, the data in this study provide a starting place for decrypting gene expression in the wild type perinatal small intestine. It is clear, however, that these patterns of gene expression are labile; they are both time-sensitive and subject to change in disease or

injury. Nevertheless, these data will be of great value for tracing signaling crosstalk and may eventually lead to the development of molecular tools for the manipulation of these signaling pathways. Functionally, this global view of the different mesenchymal and epithelial transcriptomes reveals compartmentalization of many of the molecules that direct epithelial/mesenchymal crosstalk. The data indicate that a complex but non-intestine-specific mesenchymal tissue secretes multiple soluble molecules that serve to support and direct an epithelial signature that is uniquely intestinal.

ACKNOWLEDGEMENTS

The authors acknowledge the excellent service of the Microarray Core Facility in the University of Michigan Diabetes Center. Tissue processing was facilitated by the Center for Organogenesis Morphology Core. D.L.G. acknowledges support from NIH P01-DK062041. W.Z. and Å.K. are both Fellows in the Organogenesis Training Program (NIH-T32-HL07505). The authors are grateful to Dr. Linda Samuelson for helpful discussions and critical reading of the manuscript. B.B.M. is now at University of Pennsylvania, Department of Genetics, Philadelphia, PA.

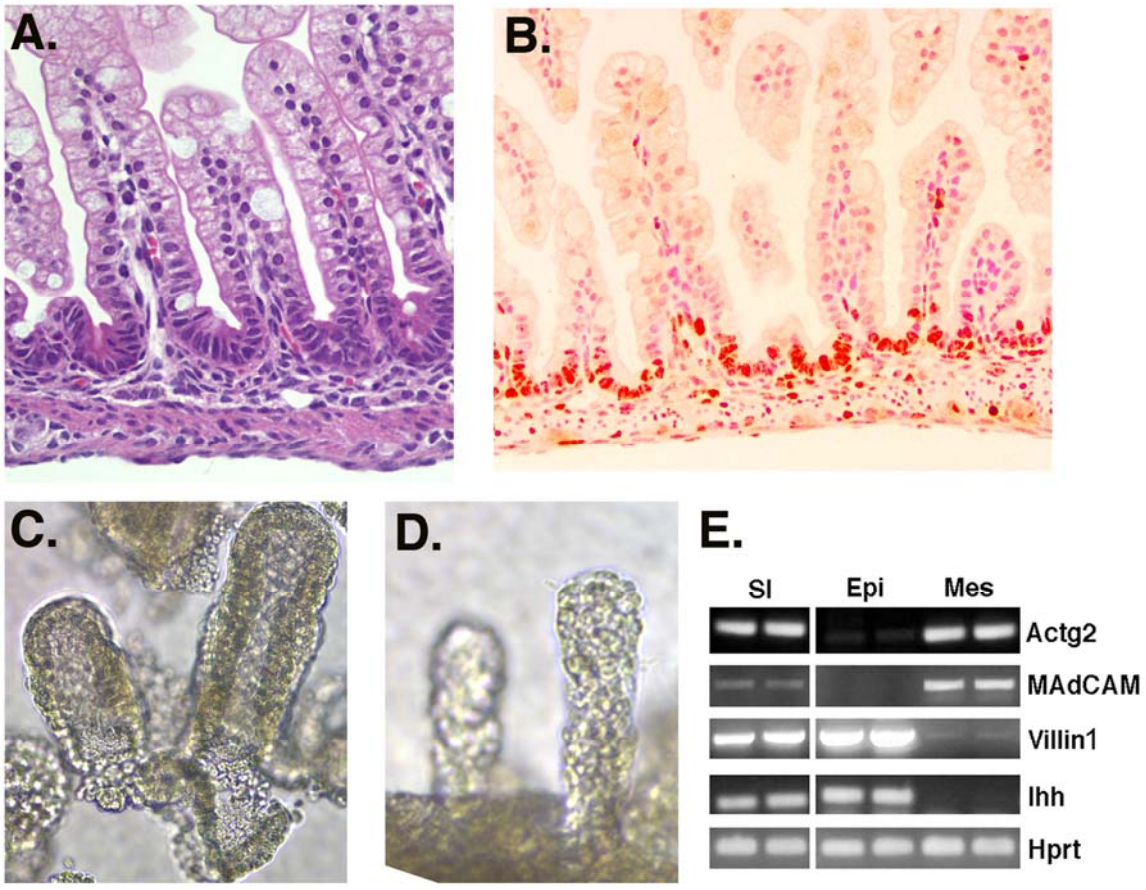

Figure 2.1 Appearance of the E18.5 intestine and characterization of separated epithelium and mesenchyme. A) This section of proximal small intestine, stained with H&E, reveals well formed villi containing absorptive cells and goblet cells (arrows). B) Staining with anti-Ki67 (dark red nuclei at bottom of villi) reveals polarization of the villi, with proliferative cells located at the villus base. Crypts have not yet formed. C) Epithelial fraction after separation of mesenchyme. D) Mesenchymal fraction after separation of epithelium. E) PCR verification of tissue separation. Primers and conditions for PCR are listed in Table 2.1. Smooth muscle gamma actin (Actg2) and MAdCAM are expressed in mesenchyme (Mes), while Villin1 and Ihh are expressed in epithelium (Epi). Hprt is expressed in all fractions. SI = unseparated small intestine.

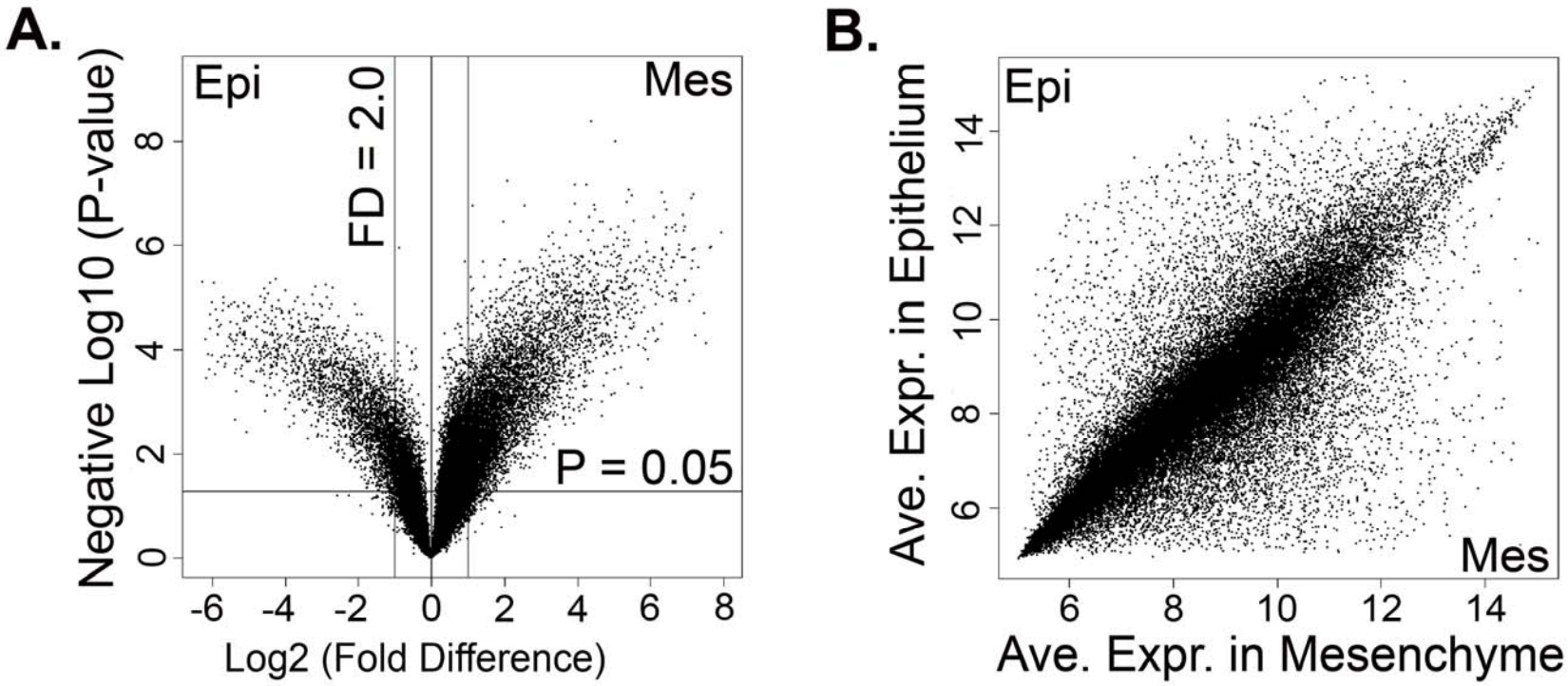

Figure 2.2 Microarray data distribution. A) As detailed in the text, the limits of FC = 2.0 and P value ≤ 0.05 demarcate the genes enriched in epithelium (Epi) and mesenchyme (Mes). B) The average expression values as calculated by RMA (described in Materials and Methods) for all probes are displayed for epithelium (y axis) and mesenchyme (x axis). Points above the diagonal are epithelially enriched genes; those below the diagonal are enriched in mesenchyme.

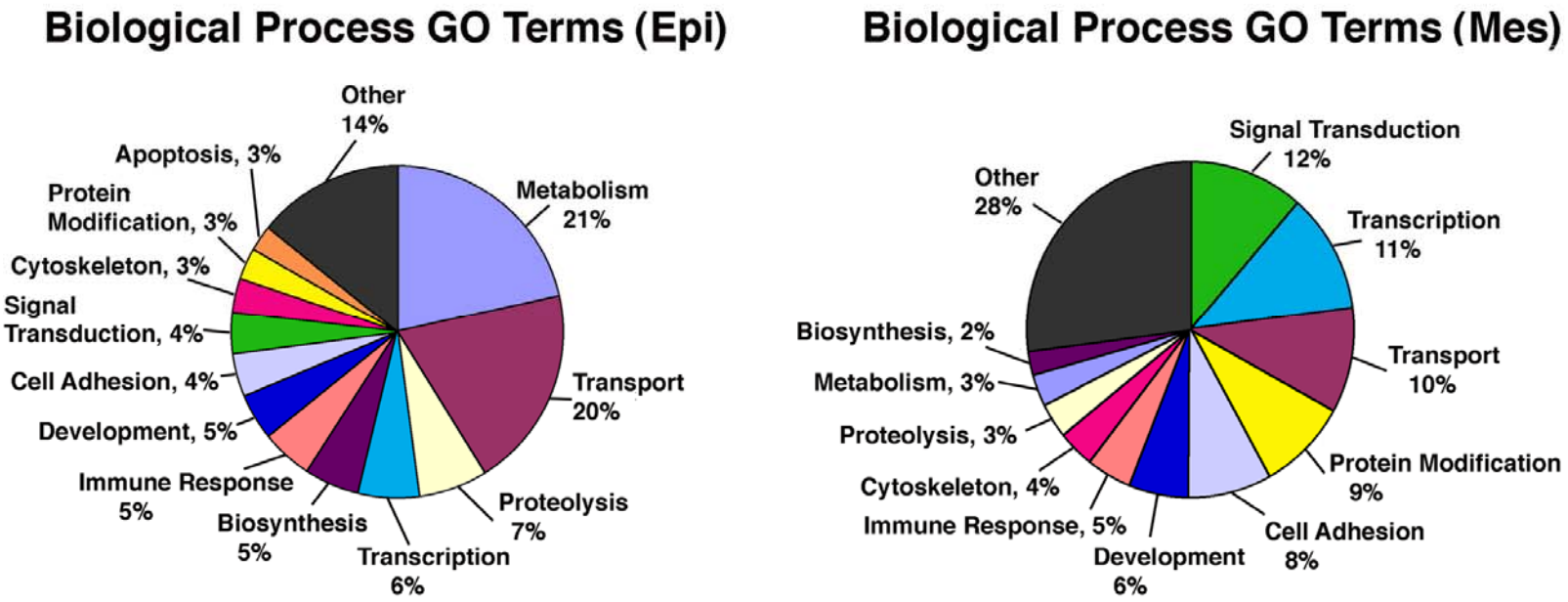

Figure 2.3 Biological Process Gene Ontology terms for epithelium (left) and mesenchyme (right). Each term is represented by a different color; the same term is represented by the same color in each compartment (i.e., green is signal transduction). The percent representation of each term is denoted schematically by the pie chart and provided numerically. All terms shared by less than one percent of the genes are grouped into the category called Other.

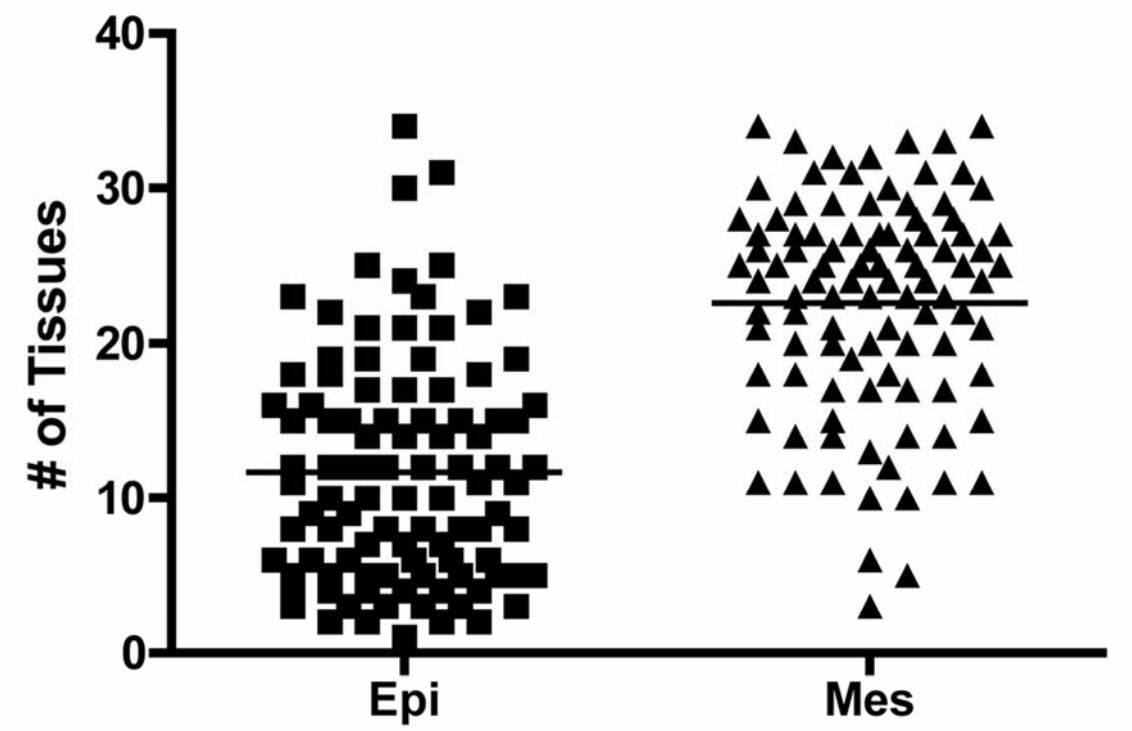

Figure 2.4 Tissue specificity of genes enriched in epithelium and mesenchyme. The number of tissues in which the top 100 epithelially and mesenchymally enriched genes are expressed was determined from EST counts using the Unigene Expression Profile viewer at NCBI, linked to the Unigene page. Each square (Epi) or triangle (Mes) indicates a separate gene. While 50% of epithelial genes are expressed in 10 or fewer tissues, only 5 of the mesenchymally enriched genes show this degree of tissue restriction.

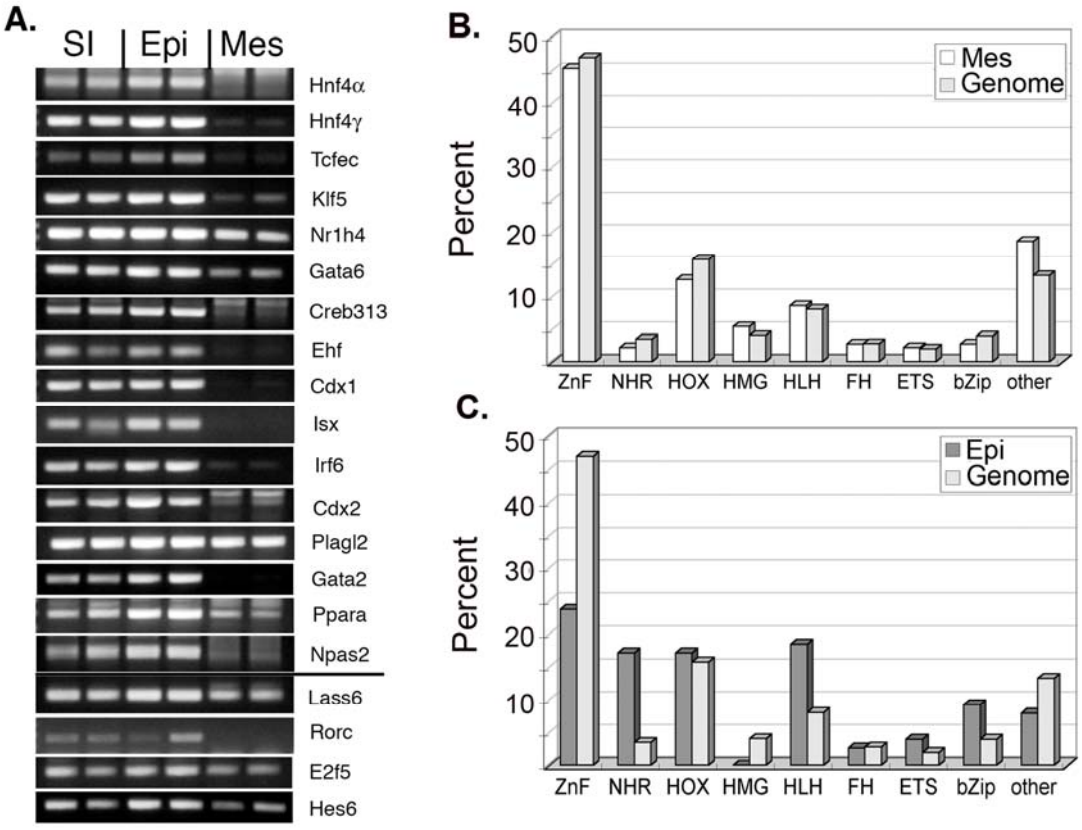

Figure 2.5 Transcription factors enriched in the epithelium. A) PCR verification of epithelial enrichment for the top 16 most enriched genes in the epithelium (above the line under Npas2). A few genes of lower enrichment values were also tested (Lass6, Rorc, E2F5 and Hes6). SI = unseparated small intestine; Epi = epithelium; Mes = mesenchyme. B) Distribution of the classes of transcription factors found in mesenchyme (white bars), compared to the distribution of classes in the entire genome (crosshatched bars). Data for the genomic distribution is from Gray et al.(Gray et al 2004). C) Distribution of the classes of transcription factors found in epithelium (dark gray bars), compared to the distribution of classes in the entire genome (crosshatched bars). ZnF = zinc finger; HOX = homeodomain; HLH = helix loop helix; HMG = High mobility group; NHR = Nuclear hormone receptor; FH = forkhead; ETS = Ets factor; Other = all other classes.

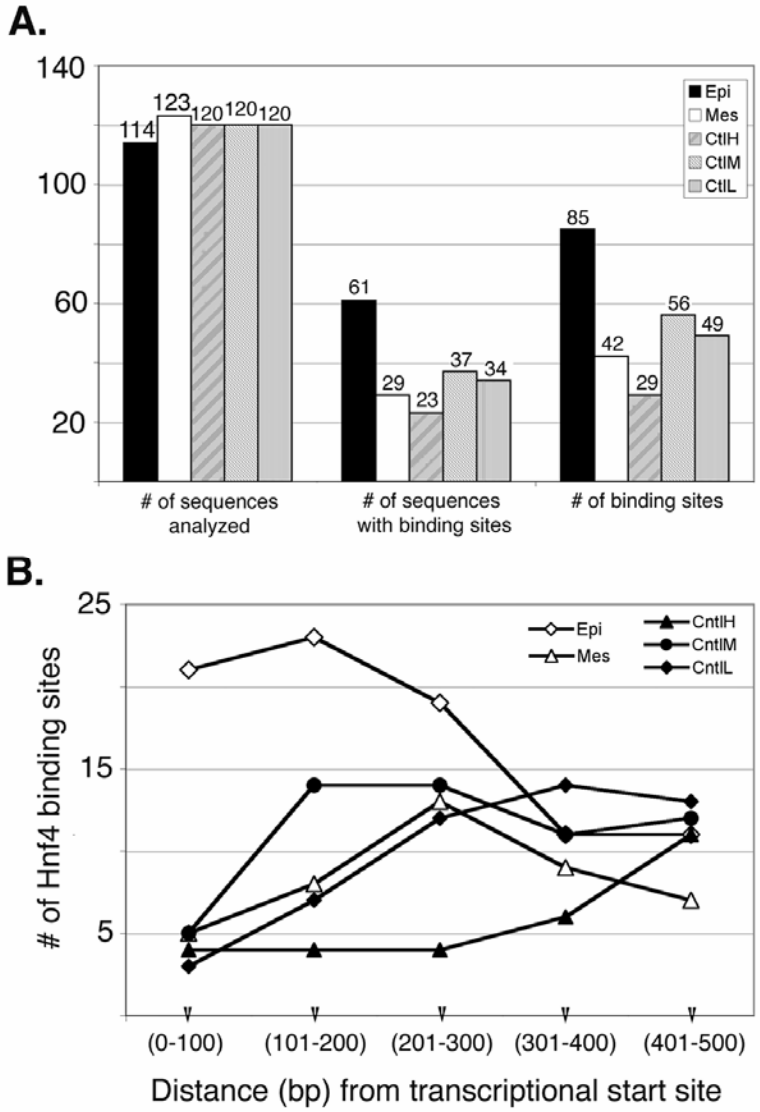

Figure 2.6 Distribution of Hnf4 binding sites. A) The total number of sequences analyzed for each group is shown in the first set of bars. Epithelium and mesenchyme are represented by the black and white bars, respectively. Three control groups of genes with ES = 1.0 are represented by various shades of grey patterned bars. The three control groups are composed of genes expressed at low (L; average expression value < 8.0), medium (M; average expression value = 8.0-10.0) and high (H; average expression value >10.0) relative expression values, respectively. The central set of bars display the number of sequences that were positive for Hnf4 binding sites. Some of these sequences have more than one binding site. The third set of bars indicates the total number of Hnf4 binding sites found in all sequences of each type searched. B) The location of Hnf4 binding sites is graphically displayed for each of five 100 bp "bins" across the 500 bp sequences analyzed. The epithelium and mesenchyme are represented by open diamonds and open triangles, respectively. The three control groups are represented by closed symbols. Note the dramatic difference between the epithelial group and all other groups in the first 100 bp.

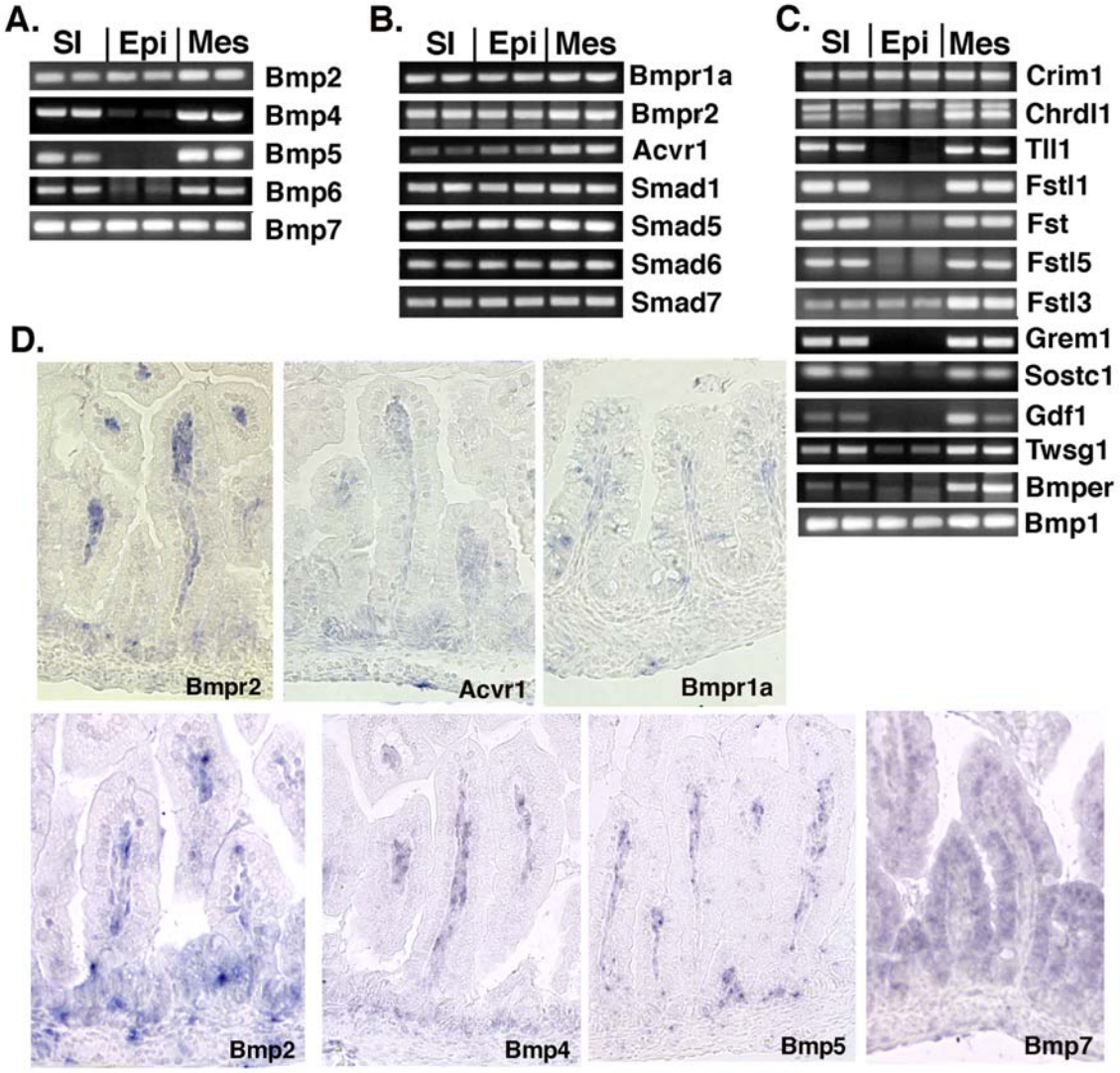

Figure 2.7 Distribution of Bmp signaling pathway molecules in epithelium and mesenchyme. A) PCR verification of microarray results for several Bmp ligands. Note that Bmp7 is the only ligand that is slightly enriched in the epithelium. B) PCR verification of array results for Bmp receptors and Smad signaling proteins. All are expressed in both epithelium and mesenchyme, with slight enrichment visible for some in the mesenchyme. C) PCR verification of the location of inhibitors and modulators of Bmp signaling. The majority of these proteins are expressed exclusively or predominantly in the mesenchyme. None are preferentially enriched in the epithelium. D) In situ hybridization on sections of intestine from E17.5 embryos, using antisense probes for various Bmp pathway molecules: Bmpr2, Acvr1, Bmpr1a, Bmp2, Bmp4, Bmp5, and Bmp7. All tissue sections are from the proximal small intestine, except for Bmpr1a, which is from the distal small intestine.

Table 2.1 PCR primers

GENE	FORWARD PRIMER	REVERSE PRIMER	Length (bp)	DMSO	$MgCl_2$ (μM)	TchDn	Cycle #
Acvr1	GGAACGAGGACCACTGTGAAGG	AACACCACCGAGAGGATGATAAGG	271	YES	1.5	NO	30
Bmp1	GAGGAAGGCTATGGCGTGGAG	TTGGAGATGGTGTCGTCAGAGTG	209	YES	1.5	NO	30
Bmp2	GAGGAGGCGAAGAAAAGCAACAG	GAAGCAGCAACACTAGAAGACAGC	178	YES	1.5	NO	30
Bmp4	AGGAGGAGGAGGAAGAGCAGAG	GTTTATACGGTGGAAGCCCTGTTC	263	NO	1.5	NO	30
Bmp5	TCTGCACTTACCACCAAGGACTC	TCTGCTGCTGTCACTGCTTCTC	213	YES	1.5	NO	30
Bmp6	TCTTCAGACTACAACGGCAGTGAG	TCGGGATTCATAAGGTGGACCAAG	215	YES	2	YES	32
Bmp7	AGGAGGGCTGGTTGGTGTTTG	TGGCGTCTTGGAGCGATTCTG	263	YES	1.5	NO	30
Bmper	AGGTGACGATGCTCTGGTTCTTC	TGACAGCACAGGGCACTTCTC	272	NO	1.5	NO	30
Bmpr1a	CCGCTATGGAGAAGTATGGATGGG	CAGACCACAAGCAGCAGAATAAGC	298	YES	1.5	NO	30
Bmpr2	CTGTGAACCTGAGGGACTGTGAG	AACTTGGGTCTCTGCTTCTCTCTG	214	YES	2	NO	30
Cdx1	CGGACGCCCTACGAATGGATG	CTGCTGCTGCTGCTGTTTCTTC	276	YES	1.5	NO	30
Cdx2	GAAACCTGTGCGAGTGGATGC	TTGTTGCTGCTGCTGCTGTTG	284	YES	1.5	NO	30
Chrdl1	GCTTTAGTCCACCGCTCCTCTC	GCAGTTCACACAGTAAACCAGTCC	256	YES	2.5	NO	30
Creb3l3	CAGTGGCATCTCTGAGGATCTACC	CAGTGAGGTTGAAGCGGGAGG	266	YES	1.5	NO	30
Crim1	ATCCCATAATGAGAGCGTGCCTAG	AATATCCCACCGTTCCTCGTCAG	274	NO	3	NO	30
E2f5	ACGGCGTCCTGGATCTCAAAG	AGTATTACAGCCAGCACCTACACC	158	NO	1.5	NO	30
Ehf	GCTTCCTGCCTTCTTCTTCATCAC	GGTTGTTGGCTGGGTTGAGATTC	119	NO	1.5	NO	30
Fst	CTGCTCTTCTGGCGTGCTTCTTG	TGTAGTCCTGGTCTTCCTCCTCCTC	101	YES	1.5	NO	30
Fstl1	CAATCGCTGTGTCTGTTCCTGTG	TCCTCCTCTGTGTGGGTCTGG	96	YES	1.5	NO	30
Fstl3	TTACCTACATCTCGTCGTGTCACC	AAATCGGGATGGCGTCAAATGC	172	YES	1.5	NO	30
Fstl5	CCCAAAGCAGAAGAGGATGAAGTG	TGAGTGTGGATGGTGTGGTGTG	296	NO	1.5	YES	32
Fzd6	TGTTGGTATCTCTGCGGTCTTCTG	GGTCGCTCCTGTGCTAGTTCC	241	YES	1.5	NO	30
Gata4	AGCAGCAGCAGCAGTGAAGAG	CGAGCAGGAATTTGAAGAGGGAAG	300	NO	1.5	NO	30
Gata6	GCTTGCGGGCTCTATATGAAACTC	TGAGGTGGTCGCTTGTGTAGAAG	219	YES	1.5	NO	30
Gdf10	AGAAGGACCAGGACACATTCACC	AGCACAGTAGTAGGCGTCAAAGG	197	NO	1.5	YES	32
Gmcl1	GGTCGCATTTGGATCACTGTATCG	TGTCTCACCGCACTGCTGTATC	121	NO	2	NO	30
Grem1	AGAATGAATCGCACCGCATACAC	GACTCAAGCACCTCCTCTCCAG	230	YES	1.5	NO	30
Hes6	CGGATCAACGAGAGTCTTCAGGAG	TTCTAGCAGGTGGTTCAGGAGTTC	276	NO	1.5	NO	30
Hnf4a	GATGCTTCTCGGAGGGTCTGC	TTGGTGGTGATGGCTGTGGAG	200	YES	1.5	NO	30

Hnf4g	CGCAGCATTCGGAAGAGTCATG	CCGCTTGTGCCAGAGTGTTTATG	220	YES	1.5	NO	30
Isx	ACTTCACCCATTACCCTGACATCC	TCTTCTCCTGCTTCCTCCACTTG	123	YES	1.5	NO	30
Mlx	CTCAGCAAAGCCATCGTTCTACAG	AGGCGTTGAAAGACTGGAAGAGG	256	YES	1.5	NO	30
Npas2	CAGTGGTCAGTTACGCAGATGTTC	GGTTGGAGTGGAGGTGGGTTC	248	YES	3	NO	30
Nr1h4	GGCATCTATGAACTCAGGCGAATG	CGGCGTTCTTGGTAATGCTTCTTC	232	YES	1.5	NO	30
Per2	CTTCCAGTCCAGAGGCAGTAGTC	CCAGAAGGAGGTTGAGCAGGTC	245	YES	1.5	YES	32
Plagl2	GGGAGGCAGAGAGTCAAGTGAAG	TGGTTCCGCAGATGGTCCTTTC	253	YES	1.5	NO	30
Ppara	GCGGGAAAGACCAGCAACAAC	TCAAGGAGGACAGCATCGTGAAG	283	YES	1.5	NO	30
Rorc	CTGCGACTGGAGGACCTTCTAC	ACTGTGTGGTTGTTGGCATTGTAG	269	NO	2.5	NO	30
Smad1	CCTGCTTACCTGCCTCCTGAAG	GCGGTTCTTATTGTTGGACGGATC	252	NO	1.5	NO	30
Smad5	AGATGTTCAGCCTGTCGCCTATG	ACTAAGACACTCAGCGTACACCTC	274	YES	1.5	NO	30
Smad6	CTATTCTCGGCTGTCTCCTCCTG	CCTGGTCGTACACCGCATAGAG	224	YES	1.5	NO	30
Smad7	AGGCTGTGTTGCTGTGAATCTTAC	CGAGTCTTCTCCTCCCAGTATGC	290	YES	3	NO	30
Sostdc1	TGGAGGCAGGCATTTCAGTAGC	AATGTATTTGGTGGACCGCAGTTC	88	NO	2.5	YES	32
Tcf23	AGGAAGAGGAGTCGCATCAACAG	CAGCACGAGCACATCCAACTTG	267	YES	2.5	YES	32
Tcfec	ATGAACCCATGAGCCCAGACAG	AGCATCCGTGAGACCAGCATTAG	173	YES	2	NO	30
Tgfbr2	CTCACCTACCACGGCTTCACTC	TGGACACGGTAGCAGTAGAAGATG	260	NO	3	NO	30
Tgfbr3	CGACTTGCCACACTTGCCATC	ACAGCCAGACAGAACGGTGAAG	212	YES	1.5	NO	30
Tll1	AAGATGGAGCCTGGAGAAGTGAAC	TGTGTAGGAAGGGTAGCCATTTGG	288	YES	1.5	YES	32
Twsg1	ACTCTGTGCCAGCGATGTGAG	AGGAGACGATGTTCCAGTTCAGC	275	NO	1.5	NO	30
HPRT	AGTCCCAGCGTCGTGATTAGC	ATAGCCCCCCTTGAGCACACAG	204	YES	1.0	NO	30
Actg2	GGGTGTGATGGTGGGAATGG	GGTGCTCTTCTGGTGCTACTC	182	YES	1.25	NO	30
MadCAM	TGCCAATCCATAGGACGACG	GCACACCACTGTTCCGAATGAC	591	YES	1.5	NO	30
Ihh	CCATCTTCATCCCAGCCTTCG	CACCCCCAACTACAATCCCG	168	YES	1.5	NO	30
Vil1	CAGTCACCACCGTAGAAATCGC	AGATGGATGACTACCTGAAGGGC	1027	YES	1.5	NO	30

DMSO = Addition of DMSO (final concentration 2%) is indicated
TchDn = Touchdown mode. Number of cycles is listed in the last column

Table 2.2 Top 100 Genes enriched in epithelium.

ProbeSet ID	Symbol	Gene Title	Epi[a]	Mes[b]	p-value	Es[c]	#Tis[d]
1421489_a_at	2010106E10Rik	RIKEN cDNA 2010106E10 gene	12.21	5.95	5.2E-06	76.5	3
1424673_at	Clec2h	C-type lectin domain family 2, member h	13.43	7.26	6.2E-05	72.0	5
1416306_at	Clca3	chloride channel calcium activated 3	12.44	6.30	3.5E-04	70.6	8
1455431_at	Slc5a1	solute carrier family 5 (sodium/glucose cotransporter), member 1	12.11	5.99	1.2E-04	69.1	10
1418734_at	H2-K1	Histocompatibility 2, K1, K region	12.83	6.75	2.5E-04	67.7	34
1450109_s_at	Abcc2	ATP-binding cassette, sub-family C (CFTR/MRP), member 2	12.19	6.14	2.1E-05	66.2	7
1427000_at	Hnf4a	hepatic nuclear factor 4, alpha	11.98	5.94	4.0E-05	65.7	9
1423439_at	Pck1	phosphoenolpyruvate carboxykinase 1, cytosolic	11.82	5.82	1.3E-04	64.3	15
1419622_at	Ugt2b5	UDP glucuronosyltransferase 2 family, polypeptide B5	11.82	5.82	1.3E-05	64.0	9
1459737_s_at	Ttr	transthyretin	12.46	6.51	5.3E-06	61.6	21
1417946_at	Abhd3	abhydrolase domain containing 3	12.70	6.78	3.0E-04	60.7	14
1448766_at	Gjb1	gap junction membrane channel protein beta 1	12.10	6.21	1.7E-05	59.2	15
1424357_at	Tmem45b	transmembrane protein 45b	12.88	7.02	2.7E-05	58.2	10
1450631_x_at	Defcr12	defensin related cryptdin 12	11.76	5.94	1.8E-04	56.8	5
1449036_at	Rnf128	ring finger protein 128	12.58	6.77	3.6E-05	56.2	22
1455961_at	Mme	Membrane metallo endopeptidase	12.90	7.14	3.4E-05	54.4	19
1419386_at	Muc13	mucin 13, epithelial transmembrane	12.77	7.03	1.4E-04	53.5	7
1456777_at	Mgam	maltase-glucoamylase	12.08	6.37	1.8E-05	52.5	8
1429467_s_at	Slc26a3	solute carrier family 26, member 3	12.39	6.70	1.3E-05	51.5	6
1435162_at	Prkg2	Protein kinase, cGMP-dependent, type II	12.78	7.11	1.0E-05	50.9	17
1448261_at	Cdh1	cadherin 1	13.21	7.55	2.2E-04	50.4	21
1460127_at	Hnf4g	Hepatocyte nuclear factor 4, gamma	11.04	5.40	1.3E-04	49.8	5
1419523_at	Cyp3a13	cytochrome P450, family 3, subfamily a, polypeptide 13	12.38	6.75	1.9E-04	49.5	6
1460606_at	Hsd17b13	hydroxysteroid (17-beta) dehydrogenase 13	11.27	5.70	8.3E-05	47.6	6
1420553_x_at	Serpina1a	serine (or cysteine) peptidase inhibitor, clade A,	11.23	5.65	8.3E-05	47.5	15

		member 1a					
1439796_at	Clca6	chloride channel calcium activated 6	13.89	8.35	1.5E-05	46.5	3
1448973_at	Sult1d1	sulfotransferase family 1D, member 1	12.92	7.41	1.4E-04	45.5	7
1436615_a_at	Otc	ornithine transcarbamylase	11.06	5.55	5.9E-05	45.4	3
1425102_a_at	Ace2	angiotensin I converting enzyme (peptidyl-dipeptidase A) 2	13.85	8.36	5.2E-05	45.0	12
1433579_at	Tmem30b	transmembrane protein 30B	12.46	7.00	6.4E-05	44.0	17
1438359_at	2010003K15Rik	RIKEN cDNA 2010003K15 gene	11.64	6.18	1.4E-05	44.0	4
1416646_at	Afp	alpha fetoprotein	13.00	7.59	6.7E-05	42.6	12
1448741_at	Slc3a1	solute carrier family 3, member 1	13.27	7.88	1.0E-04	41.8	24
1425079_at	Tm6sf2	transmembrane 6 superfamily member 2	12.88	7.51	1.3E-05	41.5	9
1448783_at	Slc7a9	solute carrier family 7 (cationic amino acid transporter, y+ system), member 9	13.49	8.14	2.9E-04	41.0	5
1418215_at	Mep1b	meprin 1 beta	13.76	8.40	7.6E-05	40.9	2
1437060_at	Olfm4	olfactomedin 4	12.47	7.12	2.4E-04	40.6	12
1450389_s_at	Pip5k1a	phosphatidylinositol-4-phosphate 5-kinase, type 1 alpha	11.01	5.68	6.6E-05	40.1	30
1427480_at	Leap2	liver-expressed antimicrobial peptide 2	13.02	7.70	5.2E-05	39.9	3
1434628_a_at	Rhpn2	rhophilin, Rho GTPase binding protein 2	12.27	6.95	1.4E-04	39.9	23
1427221_at	Xtrp3s1	X transporter protein 3 similar 1 gene	11.34	6.02	5.1E-05	39.8	14
1417797_a_at	1810019J16Rik	RIKEN cDNA 1810019J16 gene	11.02	5.72	3.2E-05	39.5	17
1447252_s_at	Mep1a	meprin 1 alpha	12.57	7.28	1.5E-05	39.1	5
1457976_at	2010002M12Rik	RIKEN cDNA 2010002M12 gene	12.93	7.65	1.3E-04	38.8	15
1437540_at	Mcoln3	mucolipin 3	12.25	7.02	3.9E-05	37.6	12
1434025_at	Klf5	Kruppel-like factor 5	12.87	7.66	2.4E-04	36.9	19
1453397_at	9130016M20Rik	RIKEN cDNA 9130016M20 gene	10.52	5.35	3.3E-04	36.0	4
1430128_a_at	Dp1l1	deleted in polyposis 1-like 1	13.14	7.97	2.1E-04	35.9	14
1438934_x_at	Sema4a	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4A	11.93	6.76	7.8E-05	35.9	31
1423411_at	BC013481	cDNA sequence BC013481	11.11	5.96	1.9E-05	35.5	23

1424959_at	Anxa13	annexin A13	13.78	8.63	1.5E-04	35.4	5
1419755_at	Mfi2	antigen p97 (melanoma associated) identified by monoclonal antibodies 133.2 and 96.5	14.20	9.06	5.2E-05	35.2	10
1452277_at	Arsg	arylsulfatase G	11.12	6.00	4.6E-05	34.9	18
1440218_at	BC040758	cDNA sequence BC040758	13.25	8.14	2.8E-05	34.6	4
1460233_at	Guca2b	guanylate cyclase activator 2b (retina)	13.31	8.20	1.3E-05	34.6	5
1449067_at	Slc2a2	solute carrier family 2 (facilitated glucose transporter), member 2	12.31	7.20	1.3E-04	34.6	5
1425298_a_at	Birc1a	baculoviral IAP repeat-containing 1a	11.64	6.53	4.9E-05	34.4	2
1433610_at	AA986860	expressed sequence AA986860	10.94	5.84	1.2E-04	34.2	15
1430637_at	2210016H18Rik	RIKEN cDNA 2210016H18 gene	11.62	6.53	5.4E-05	34.2	4
1455454_at	Akr1c19	aldo-keto reductase family 1, member C19	13.67	8.57	4.4E-05	34.2	8
1450060_at	Pigr	polymeric immunoglobulin receptor	12.14	7.04	2.5E-05	34.2	8
1427137_at	Ces5	carboxylesterase 5	12.07	6.99	6.0E-05	33.9	12
1455593_at	Apob	apolipoprotein B	13.29	8.25	1.7E-04	32.9	8
1426911_at	Dsc2	desmocollin 2	10.81	5.78	2.8E-05	32.8	18
1428595_at	Slc6a19	solute carrier family 6 (neurotransmitter transporter), member 19	10.77	5.75	2.3E-05	32.5	12
1424649_a_at	Tspan8	tetraspanin 8	12.71	7.69	1.7E-04	32.4	16
1426261_s_at	Ugt1a2	UDP glucuronosyltransferase 1 family, polypeptide A2	12.92	7.92	1.5E-05	32.0	23
1457658_x_at	Anxa4	annexin A4	12.02	7.04	2.4E-05	31.4	25
1427961_s_at	Ugt2b34	UDP glucuronosyltransferase 2 family, polypeptide B34	12.42	7.45	1.6E-04	31.3	5
1423858_a_at	Hmgcs2	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2	13.35	8.41	4.0E-04	30.7	22
1438994_at	Hyal5	hyaluronoglucosaminidase 5	12.25	7.31	2.3E-05	30.6	3
1418940_at	Sult1b1	sulfotransferase family 1B, member 1	12.59	7.68	6.3E-05	30.1	4
1419294_at	1700011H14Rik	RIKEN cDNA 1700011H14 gene	11.67	6.76	1.4E-04	30.0	3
1417089_a_at	Ckmt1	creatine kinase, mitochondrial 1, ubiquitous	13.93	9.05	8.5E-05	29.4	14
1418672_at	Akr1c13	aldo-keto reductase family 1, member C13	13.43	8.57	4.3E-04	29.0	15
1438980_x_at	4732466D17Rik	RIKEN cDNA 4732466D17 gene	10.47	5.64	6.9E-05	28.3	2

1428357_at	2610019F03Rik	RIKEN cDNA 2610019F03 gene	13.27	8.46	7.2E-05	28.0	15
1421040_a_at	Gsta2	glutathione S-transferase, alpha 2 (Yc2)	11.86	7.07	2.6E-04	27.6	10
1426990_at	Cubn	cubilin (intrinsic factor-cobalamin receptor)	12.90	8.12	7.9E-05	27.5	11
1448873_at	Ocln	occludin	11.12	6.34	1.5E-04	27.5	15
1457555_at	Gpr151	G protein-coupled receptor 151	11.15	6.37	9.1E-05	27.4	2
1439260_a_at	Enpp3	Ectonucleotide pyrophosphatase/phosphodiesterase 3	13.11	8.33	3.6E-05	27.4	4
1452298_a_at	Myo5b	myosin Vb	12.34	7.57	6.3E-05	27.2	16
1415938_at	Spink3	serine peptidase inhibitor, Kazal type 3	13.83	9.08	2.4E-05	27.0	8
1416271_at	Perp	PERP, TP53 apoptosis effector	12.15	7.44	1.1E-04	26.2	19
1427042_at	Mal2	mal, T-cell differentiation protein 2	10.28	5.58	1.1E-04	26.1	12
1455901_at	Chpt1	choline phosphotransferase 1	12.72	8.03	1.7E-04	25.8	25
1449091_at	Cldn8	claudin 8	10.81	6.12	2.7E-04	25.7	12
1448837_at	Vil1	villin 1	13.33	8.65	1.5E-04	25.7	16
1448964_at	S100g	S100 calcium binding protein G	11.37	6.70	1.7E-05	25.5	6
1434915_s_at	Lrrc19	leucine rich repeat containing 19	10.09	5.41	1.2E-04	25.5	6
1450947_at	2610528J11Rik	RIKEN cDNA 2610528J11 gene	11.37	6.71	1.1E-04	25.3	11
1418777_at	Ccl25	chemokine (C-C motif) ligand 25	12.53	7.87	2.3E-04	25.3	16
1428663_at	5133401H06Rik	RIKEN cDNA 5133401H06 gene	10.96	6.32	1.3E-04	25.0	11
1416579_a_at	Tacstd1	tumor-associated calcium signal transducer 1	14.13	9.50	4.3E-05	24.7	19
1439479_at	Lct	lactase	14.28	9.67	2.5E-04	24.3	6
1422868_s_at	Gda	guanine deaminase	10.97	6.38	9.1E-05	24.0	18
1418654_at	Hao3	hydroxyacid oxidase (glycolate oxidase) 3	11.42	6.84	1.5E-04	23.9	8
1448470_at	Fbp1	fructose bisphosphatase 1	13.68	9.10	6.9E-06	23.9	12
1448382_at	Ehhadh	enoyl-Coenzyme A, hydratase/3-hydroxyacyl Coenzyme A dehydrogenase	11.46	6.88	2.0E-04	23.8	21

[a]Epi = Average expression value in Epithelium (Log2) as calculated by RMA.
[b]Mes = Average expression value in Mesenchyme (Log2) as calculated by RMA.
[c]ES = Enrichment Score (absolute value of numerical fold difference, calculated as in Materials and Methods).
[d]#Tis = For the first 100 genes, the number of tissues in which the gene is expressed (as determined from the NCBI Unigene Expression Profile) is provided.

Table 2.3 Top 100 Genes enriched in Mesenchyme.

Probe Set ID	Symbol	Gene Title	Epi[a]	Mes[b]	pValue	ES[c]	#Tis[d]
1456292_a_at	Vim	vimentin	5.39	13.34	5.7E-07	246.4	34
1451263_a_at	Fabp4	fatty acid binding protein 4, adipocyte	5.63	13.28	1.1E-06	201.5	25
1424769_s_at	Cald1	caldesmon 1	6.01	13.63	1.1E-06	196.3	33
1452670_at	Myl9	myosin, light polypeptide 9, regulatory	7.01	14.53	7.5E-05	184.2	24
1424114_s_at	Lamb1-1	laminin B1 subunit 1	5.92	13.22	2.7E-06	158.1	26
1422340_a_at	Actg2	actin, gamma 2, smooth muscle, enteric	7.13	14.37	5.4E-05	150.6	13
1427883_a_at	Col3a1	procollagen, type III, alpha 1	7.34	14.56	1.2E-06	148.5	33
1460208_at	Fbn1	fibrillin 1	5.65	12.83	1.1E-07	144.3	20
1423669_at	Col1a1	procollagen, type I, alpha 1	6.96	14.13	8.8E-06	144.2	27
1437992_x_at	Gja1	gap junction membrane channel protein alpha 1	5.51	12.66	1.2E-06	141.5	30
1459749_s_at	Fat4	FAT tumor suppressor homolog 4	5.12	12.26	2.9E-06	140.9	23
1416514_a_at	Fscn1	fascin homolog 1, actin bundling protein (Strongylocentrotus) purpuratus)	5.38	12.52	1.3E-07	140.6	31
1421917_at	Pdgfra	platelet derived growth factor receptor, alpha polypeptide	6.30	13.25	4.9E-06	124.0	24
1450757_at	Cdh11	cadherin 11	5.74	12.69	4.8E-06	123.5	27
1448254_at	Ptn	pleiotrophin	5.33	12.27	1.7E-06	123.3	28
1448789_at	Aldh1a3	aldehyde dehydrogenase family 1, subfamily A3	5.47	12.41	4.3E-07	122.5	14
1452106_at	Npnt	nephronectin	5.40	12.32	5.2E-07	121.4	18
1450843_a_at	Serpinh1	serine (or cysteine) peptidase inhibitor, clade H, member 1	6.67	13.55	2.7E-07	117.7	29
1425028_a_at	Tpm2	tropomyosin 2, beta	7.34	14.19	1.7E-06	115.7	24
1450852_s_at	F2r	coagulation factor II (thrombin) receptor	5.81	12.59	9.8E-07	110.4	25
1437726_x_at	C1qb	complement component 1, q subcomponent, beta polypeptide	5.32	12.09	1.2E-06	109.4	31
1456380_x_at	Cnn3	calponin 3, acidic	5.41	12.14	5.9E-06	105.8	32
1448421_s_at	Aspn	asporin	5.09	11.82	1.5E-06	105.8	15

1452163_at	Ets1	E26 avian leukemia oncogene 1, 5' domain	5.55	12.25	5.5E-07	104.3	25
1456739_x_at	Armcx2	armadillo repeat containing, X-linked 2	6.12	12.82	1.2E-05	104.0	27
1437347_at	Ednrb	endothelin receptor type B	6.80	13.48	8.7E-06	102.6	23
1423062_at	Igfbp3	insulin-like growth factor binding protein 3	5.72	12.40	4.1E-07	102.3	26
1417789_at	Ccl11	small chemokine (C-C motif) ligand 11	5.86	12.50	2.0E-06	99.3	11
1448943_at	Nrp1	neuropilin 1	5.91	12.52	7.7E-06	97.7	26
1448962_at	Myh11	myosin, heavy polypeptide 11, smooth muscle	7.19	13.77	4.7E-07	96.2	15
1416779_at	Sdpr	serum deprivation response	5.56	12.14	2.8E-06	95.6	23
1448392_at	Sparc	secreted acidic cysteine rich glycoprotein	7.31	13.82	8.5E-06	91.7	34
1418726_a_at	Tnnt2	troponin T2, cardiac	5.28	11.73	4.6E-07	87.6	11
1423607_at	Lum	lumican	6.49	12.94	2.1E-07	87.1	27
1417466_at	Rgs5	regulator of G-protein signaling 5	6.60	13.05	1.5E-06	86.9	26
1417447_at	Tcf21	transcription factor 21	6.30	12.72	5.0E-06	85.3	14
1420872_at	Gucy1b3	guanylate cyclase 1, soluble, beta 3	5.70	12.10	1.5E-07	84.2	23
1429159_at	4631408O11 Rik	RIKEN cDNA 4631408O11 gene	6.11	12.45	9.9E-06	80.6	14
1424443_at	Tm6sf1	transmembrane 6 superfamily member 1	5.27	11.61	1.2E-06	80.6	22
1436970_a_at	Pdgfrb	platelet derived growth factor receptor, beta polypeptide	5.50	11.82	1.1E-06	80.4	25
1447862_x_at	Thbs2	thrombospondin 2	5.22	11.54	1.0E-07	80.0	21
1416221_at	Fstl1	follistatin-like 1	6.90	13.21	9.8E-08	79.3	30
1435383_x_at	Ndn	necdin	6.00	12.31	2.8E-06	79.3	28
1438651_a_at	Agtrl1	angiotensin receptor-like 1	6.49	12.78	3.2E-05	78.7	24
1460187_at	Sfrp1	secreted frizzled-related sequence protein 1	6.00	12.29	3.9E-06	78.5	25
1434141_at	Gucy1a3	guanylate cyclase 1, soluble, alpha 3	5.57	11.85	2.4E-06	78.0	18
1455439_a_at	Lgals1	lectin, galactose binding, soluble 1	8.13	14.33	3.4E-06	73.8	27
1432466_a_at	Apoe	apolipoprotein E	7.14	13.32	5.4E-10	72.2	33
1423608_at	Itm2a	integral membrane protein 2A	6.85	13.01	6.9E-05	71.2	22
1455907_x_at	Phox2b	paired-like homeobox 2b	5.04	11.19	1.4E-05	70.8	6
1423110_at	Col1a2	procollagen, type I, alpha 2	7.49	13.62	5.2E-06	69.9	32
1423505_at	Tagln	transgelin	8.11	14.22	1.7E-06	69.1	24
1419663_at	Ogn	osteoglycin	6.20	12.31	1.7E-06	69.0	17

1451179_a_at	Qk	quaking	6.20	12.30	4.2E-05	68.8	28
1420514_at	Tmem47	transmembrane protein 47	5.09	11.17	2.9E-07	68.0	17
1419476_at	Adamdec1	ADAM-like, decysin 1	8.31	14.38	3.8E-05	67.2	5
1436363_a_at	Nfix	nuclear factor I/X	6.28	12.28	7.3E-06	64.3	22
1457871_at	Colec10	collectin sub-family member 10	5.70	11.68	3.7E-06	63.0	3
1417327_at	Cav2	caveolin 2	5.70	11.66	2.0E-06	62.3	21
1417818_at	Wwtr1	WW domain containing transcription regulator 1	5.85	11.80	4.8E-06	61.9	27
1425810_a_at	Csrp1	cysteine and glycine-rich protein 1	7.41	13.34	1.5E-04	61.1	30
1455870_at	Akap2	A kinase (PRKA) anchor protein 2	5.94	11.82	8.6E-07	58.8	23
1449145_a_at	Cav1	caveolin, caveolae protein 1	7.85	13.72	5.9E-05	58.6	27
1416179_a_at	Rdx	radixin	6.54	12.39	1.0E-04	57.8	27
1454984_at	AW061234	expressed sequence AW061234	5.67	11.51	3.3E-06	57.4	17
1439078_at	Klhl4	kelch-like 4	5.14	10.98	1.6E-05	57.1	11
1447643_x_at	Snai2	snail homolog 2	5.97	11.79	2.7E-06	56.4	11
1449178_at	Pdlim3	PDZ and LIM domain 3	7.11	12.92	9.1E-05	56.1	18
1416740_at	Col5a1	procollagen, type V, alpha 1	7.36	13.15	6.0E-06	55.4	26
1453304_s_at	Ly6e	lymphocyte antigen 6 complex, locus E	7.01	12.77	5.4E-04	54.3	25
1423294_at	Mest	mesoderm specific transcript	7.85	13.59	1.9E-05	53.6	26
1448797_at	Elk3	ELK3, member of ETS oncogene family	6.00	11.73	3.1E-06	53.0	27
1456389_at	Zfhx1b	zinc finger homeobox 1b	5.27	10.97	1.7E-06	52.2	29
1434180_at	Plekhc1	pleckstrin homology domain containing, family C (with FERM domain) member 1	7.43	13.12	8.1E-05	51.4	29
1436448_a_at	Ptgs1	prostaglandin-endoperoxide synthase 1	5.52	11.17	1.1E-05	50.3	24
1415871_at	Tgfbi	transforming growth factor, beta induced	7.16	12.78	1.9E-05	48.9	31
1418090_at	Plvap	plasmalemma vesicle associated protein	7.20	12.81	2.0E-05	48.8	14
1427231_at	Robo1	roundabout homolog 1	5.13	10.73	8.5E-06	48.5	20
1426677_at	Flna	filamin, alpha	8.17	13.76	3.0E-06	48.2	29
1454830_at	Fbn2	fibrillin 2	6.82	12.41	6.0E-06	48.1	20
1418497_at	Fgf13	fibroblast growth factor 13	5.84	11.41	1.3E-05	47.6	12
1418664_at	Mpdz	multiple PDZ domain protein	5.55	11.11	3.3E-06	47.4	19
1416114_at	Sparcl1	SPARC-like 1 (mast9, hevin)	7.99	13.54	3.4E-07	46.9	26

1423754_at	Ifitm3	interferon induced transmembrane protein 3	7.93	13.47	3.6E-05	46.6	25
1448590_at	Col6a1	procollagen, type VI, alpha 1	8.37	13.90	8.8E-06	46.3	26
1434148_at	Tcf4	transcription factor 4	6.35	11.88	3.0E-05	46.2	29
1424807_at	Lama4	laminin, alpha 4	6.80	12.32	5.7E-06	46.0	22
1449033_at	Tnfrsf11b	tumor necrosis factor receptor superfamily, member 11b (osteoprotegerin)	5.03	10.52	4.0E-05	45.1	15
1436870_s_at	AU041783	expressed sequence AU041783	5.27	10.73	2.2E-06	44.0	21
1448471_a_at	Ctla2a	cytotoxic T lymphocyte-associated protein 2 alpha	5.11	10.56	6.5E-06	43.8	20
1427053_at	Abi3bp	ABI gene family, member 3 (NESH) binding protein	5.10	10.56	1.1E-05	43.7	21
1417917_at	Cnn1	calponin 1	7.99	13.41	6.3E-06	42.9	10
1448529_at	Thbd	thrombomodulin	6.15	11.56	8.1E-07	42.5	20
1435399_at	2310068J10 Rik	RIKEN cDNA 2310068J10 gene	6.70	12.11	8.7E-08	42.4	11
1423756_s_at	Igfbp4	insulin-like growth factor binding protein 4	8.00	13.41	9.0E-05	42.4	31
1433512_at	Fli1	Friend leukemia integration 1	5.23	10.64	7.1E-06	42.4	25
1416808_at	Nid1	nidogen 1	7.48	12.87	4.5E-04	42.0	28
1424733_at	P2ry14	purinergic receptor P2Y, G-protein coupled, 14	5.79	11.16	3.2E-06	41.5	18
1422644_at	Sh3bgr	Putative SH3BGR protein	6.18	11.55	1.0E-06	41.4	10
1425814_a_at	Calcrl	calcitonin receptor-like	5.57	10.94	1.3E-05	41.4	17

[a]Epi = Average expression value in epithelium (Log2) as calculated by RMA.
[b]Mes = Average expression value in mesenchyme (Log2) as calculated by RMA.
[c]ES = Enrichment score (absolute value of numerical fold difference, calculated as in Materials and Methods).
[d]#Tis = For the first 100 genes, the number of tissues in which the gene is expressed (as determined from the NCBI Unigene Expression Profile) is provided

Table 2.4 Array data for known compartmentalized genes.

Symbol	Gene Title	Epi[a]	Mes[b]	p-value	ES[c]	Cmpt[d]
Ppig	peptidyl-prolyl isomerase G (cyclophilin G)	10.50	10.67	5.4E-01	1.1	Both
Hprt	hypoxanthine guanine phosphoribosyl transferase 1	11.76	11.46	2.4E-01	1.2	Both
Gapdh	glyceraldehyde-3-phosphate dehydrogenase	14.21	14.09	3.6E-01	1.1	Both
Actb	actin, beta, cytoplasmic	14.60	14.61	9.3E-01	1.0	Both
Vil1	villin 1	13.33	8.65	1.5E-04	25.7	Epi
Fabp1	fatty acid binding protein 1, liver	15.09	11.33	7.9E-04	13.5	Epi
Cdx1	caudal type homeo box 1	10.51	7.26	2.0E-04	9.6	Epi
Cdx2	caudal type homeo box 2	10.90	8.03	7.4E-04	7.3	Epi
Ihh	Indian hedgehog	9.99	8.52	1.5E-03	2.8	Epi
Vim	vimentin	5.39	13.34	5.7E-07	246.4	Mes
Actg2	actin, gamma 2, smooth muscle, enteric	7.13	14.37	5.4E-05	150.6	Mes
Gli1	GLI-Kruppel family member GLI1	7.78	10.44	1.6E-04	6.3	Mes
Foxf1a	forkhead box F1a	7.85	12.52	8.1E-06	25.4	Mes
Bmp4	bone morphogenetic protein 4	7.79	12.36	5.5E-06	23.8	Mes

[a]Epi = Average expression value in epithelium (Log2) as calculated by RMA.
[b]Mes = Average expression value in mesenchyme (Log2) as calculated by RMA.
[c]ES = Enrichment score (absolute value of numerical fold difference, calculated as in Materials and Methods).
[d]Cmpt = Compartment in which the gene is expressed.

Table 2.5 Transcription factors enriched in epithelium.

Symbol	Gene Title	Epi[a]	Mes[b]	p-value	ES[c]	TF-Family[d]
Hnf4a	hepatic nuclear factor 4, alpha	11.98	5.94	4.0E-05	65.7	NHR
Hnf4g	Hepatocyte nuclear factor 4, gamma	11.04	5.40	1.3E-04	49.8	NHR
Tcfec	transcription factor EC	12.01	7.82	5.1E-05	18.3	HLH
Klf5	Kruppel-like factor 5	12.45	8.65	5.0E-05	13.9	ZnF
Nr1h4	nuclear receptor subfamily 1, group H, member 4	11.88	8.17	1.2E-04	13.0	NHR
Gata6	GATA binding protein 6	12.17	8.52	6.6E-05	12.6	ZnF
Creb3l3	cAMP responsive element binding protein 3-like 3	13.37	9.87	1.3E-05	11.3	bZip
Ehf	ets homologous factor	9.20	5.73	2.0E-04	11.1	ETS
Cdx1	caudal type homeo box 1	10.51	7.26	2.0E-04	9.6	HOX
Isx	intestine specific homeobox	11.33	8.27	2.5E-04	8.3	HOX
Irf6	interferon regulatory factor 6	11.68	8.77	5.3E-04	7.5	HLH
Cdx2	caudal type homeo box 2	10.90	8.03	7.4E-04	7.3	HOX
Plagl2	pleiomorphic adenoma gene-like 2	11.42	8.60	8.5E-04	7.1	ZnF
Gata4	GATA binding protein 4	10.66	7.84	4.4E-04	7.0	ZnF
Ppara	peroxisome proliferator activated receptor alpha	10.02	7.30	6.2E-04	6.6	NHR
Npas2	neuronal PAS domain protein 2	9.06	6.38	4.0E-05	6.4	HLH
Tcf23	transcription factor 23	9.13	6.48	9.2E-04	6.3	HLH
Nr1i3	nuclear receptor subfamily 1, group I, member 3	10.37	7.76	1.2E-04	6.1	NHR
Cebpa	CCAAT/enhancer binding protein (C/EBP), alpha	11.85	9.27	3.5E-04	6.0	bZip
Ovol1	OVO homolog-like 1	9.94	7.38	1.0E-03	5.9	ZnF
Foxa1	forkhead box A1	9.31	6.80	4.9E-04	5.7	FH
Elf3	E74-like factor 3	11.99	9.48	9.8E-05	5.7	ETS
Gata5	GATA binding protein 5	9.96	7.49	9.3E-04	5.5	ZnF
Mafb	v-maf musculoaponeurotic fibrosarcoma oncogene family, protein B	9.62	7.19	9.3E-04	5.4	bZip
Prdm16	PR domain containing 16	10.11	7.72	4.9E-04	5.2	ZnF
Esrra	Estrogen related receptor, alpha	11.55	9.17	2.5E-04	5.2	NHR

Nr2e3	nuclear receptor subfamily 2, group E, member 3	11.32	9.11	1.3E-03	4.6	NHR
Wiz	Widely-interspaced zinc finger motifs, transcript variant 2	11.08	8.93	2.8E-03	4.5	ZnF
Nr1i2	nuclear receptor subfamily 1, group I, member 2	9.48	7.37	9.9E-04	4.3	NHR
Myb	myeloblastosis oncogene	7.95	5.87	7.7E-03	4.2	other
Ipf1	insulin promoter factor 1, homeodomain transcription factor	8.92	6.95	9.7E-04	3.9	HOX
Prdm1	PR domain containing 1, with ZNF domain	11.44	9.59	3.1E-04	3.6	ZnF
Zfp54	zinc finger protein 54	8.24	6.41	1.4E-02	3.6	ZnF
Bhlhb2	basic helix-loop-helix domain containing, class B2	11.95	10.16	1.4E-03	3.4	HLH
Mxd1	MAX dimerization protein 1	10.87	9.09	3.4E-03	3.4	HLH
Rorc	RAR-related orphan receptor gamma	8.87	7.10	7.4E-05	3.4	NHR
Bach1	BTB and CNC homology 1	10.09	8.32	1.4E-02	3.4	bZip
Zbtb16	zinc finger and BTB domain containing 16	9.83	8.07	4.8E-04	3.4	ZnF
Thrb	Thyroid hormone receptor beta	8.72	6.98	4.3E-03	3.3	NHR
Tcf2	transcription factor 2	9.39	7.67	5.3E-03	3.3	HOX
Lass6	longevity assurance homolog 6	11.17	9.46	8.7E-04	3.3	HOX
Foxa3	forkhead box A3	8.89	7.22	4.9E-03	3.2	FH
Hod	homeobox only domain	10.72	9.05	7.6E-03	3.2	HOX
Lsr	liver-specific bHLH-Zip transcription factor	11.46	9.81	1.4E-03	3.1	HLH
Atoh1	atonal homolog 1	9.13	7.52	7.6E-03	3.0	HLH
Zbtb7a	zinc finger and BTB domain containing 7a	9.00	7.40	7.0E-04	3.0	ZnF
E2f8	E2F transcription factor 8	11.56	9.96	2.5E-03	3.0	other
Erf	Ets2 repressor factor	9.36	7.85	1.2E-02	2.9	ETS
Zfpm1	zinc finger protein, multitype 1	8.70	7.20	7.7E-03	2.8	ZnF
Nr5a2	nuclear receptor subfamily 5, group A, member 2	9.22	7.72	3.3E-03	2.8	NHR
Prrx1	paired related homeobox 1	8.74	7.29	7.3E-03	2.7	HOX
Pax3	paired box gene 3	8.93	7.48	3.5E-03	2.7	HOX
Srebf1	sterol regulatory element binding factor 1	11.96	10.57	2.2E-03	2.6	HLH
Nr0b2	nuclear receptor subfamily 0, group B, member 2	9.51	8.16	8.0E-03	2.6	NHR
Maf	avian musculoaponeurotic fibrosarcoma (v-maf) AS42 oncogene homolog	11.20	9.93	8.0E-04	2.4	bZip
Nkx1-2	NK1 transcription factor related, locus 2	7.80	6.54	3.1E-02	2.4	HOX

Mxi1	Max interacting protein 1	12.16	10.92	6.4E-04	2.4	HLH
Nr3c2	nuclear receptor subfamily 3, group C, member 2	7.66	6.43	1.5E-02	2.4	NHR
Neurod1	neurogenic differentiation 1	7.58	6.36	1.2E-02	2.3	HLH
Mlx	MAX-like protein X	10.38	9.18	3.4E-03	2.3	HLH
Vax2	ventral anterior homeobox containing gene 2	9.21	8.01	3.0E-03	2.3	HOX
E2f5	E2F transcription factor 5	10.65	9.46	4.2E-02	2.3	other
Gcl	germ cell-less homolog	8.41	7.22	2.1E-03	2.3	bZip
Pax6	paired box gene 6	8.05	6.86	1.4E-03	2.3	HOX
1700029I01Rik	RIKEN cDNA 1700029I01 gene	9.30	8.16	1.4E-02	2.2	ZnF
Irf7	interferon regulatory factor 7	10.85	9.73	3.4E-02	2.2	other
1700065O13Rik	RIKEN cDNA 1700065O13 gene	8.46	7.34	2.6E-02	2.2	ZnF
Hmga1	high mobility group AT-hook 1	11.78	10.69	1.8E-02	2.1	other
Hes6	hairy and enhancer of split 6	10.67	9.59	1.9E-03	2.1	HLH
Zfp36	zinc finger protein 36	11.60	10.53	8.9E-03	2.1	ZnF
Cebpg	CCAAT/enhancer binding protein (C/EBP), gamma	10.02	8.96	4.5E-03	2.1	bZip
Klf4	Kruppel-like factor 4 (gut)	9.92	8.87	2.6E-02	2.1	ZnF
Max	Max protein	9.03	8.01	5.2E-02	2.0	HLH
Zbtb7b	Vzinc finger and BTB domain containing 7B	10.92	9.90	1.4E-02	2.0	ZnF
Pitx2	paired-like homeodomain transcription factor 2	9.49	8.47	2.9E-02	2.0	HOX
Tbx15	T-box 15	8.10	7.09	7.7E-03	2.0	other

[a]Epi = Average expression value in epithelium (Log2) as calculated by RMA.
[b]Mes = Average expression value in mesenchyme (Log2) as calculated by RMA.
[c]ES = Enrichment score (absolute value of numerical fold difference, calculated as in Materials and Methods).
[d]TF-Family = Transcription factor family.

Table 2.6 Transcription factors enriched in mesenchyme.

Symbol	Gene Title	Epi[a]	Mes[b]	p-value	ES[c]	TF-Family[d]
Ets1	E26 avian leukemia oncogene 1	5.55	12.25	5.5E-07	104.3	ETS
Tcf21	transcription factor 21	6.30	12.72	5.0E-06	85.3	HLH
Ndn	necdin	6.00	12.31	2.8E-06	79.3	other
Phox2b	paired-like homeobox 2b	5.04	11.19	1.4E-05	70.8	HOX
Nfix	nuclear factor I/X	6.28	12.28	7.3E-06	64.3	other
Snai2	snail homolog 2	5.97	11.79	2.7E-06	56.4	ZnF
Elk3	ELK3, member of ETS oncogene family	6.00	11.73	3.1E-06	53.0	ETS
Zfhx1b	zinc finger homeobox 1b	5.27	10.97	1.7E-06	52.2	ZnF
Tcf4	transcription factor 4	6.35	11.88	3.0E-05	46.2	HLH
Fli1	Friend leukemia integration 1	5.23	10.64	7.1E-06	42.4	ETS
Cebpd	CCAAT/enhancer binding protein (C/EBP), delta	6.15	11.38	6.5E-06	37.5	bZip
Id4	inhibitor of DNA binding 4	6.59	11.65	3.5E-05	33.4	HLH
Pbx3	pre B-cell leukemia transcription factor 3	5.47	10.53	1.0E-08	33.2	HOX
Hoxa5	homeo box A5	6.57	11.54	8.1E-05	31.2	HOX
Sox17	SRY-box containing gene 17	5.59	10.42	2.1E-06	28.5	HMG
Trps1	Trichorhinophalangeal syndrome I	5.37	10.17	7.1E-06	27.9	ZnF
Hmgn3	high mobility group nucleosomal binding domain 3	7.16	11.93	5.2E-05	27.2	HMG
Dach1	dachshund 1	5.40	10.11	5.6E-05	26.2	other
Foxf1a	forkhead box F1a	7.85	12.52	8.1E-06	25.4	FH
Meis1	myeloid ecotropic viral integration site 1	6.13	10.78	8.2E-05	25.1	HOX
Mef2c	myocyte enhancer factor 2C	5.93	10.53	1.1E-05	24.2	other
Mrg1	myeloid ecotropic viral integration site-related gene 1	6.00	10.53	4.0E-05	23.0	other
Foxp2	forkhead box P2	5.44	9.82	5.8E-06	20.8	FH
Hoxc8	homeo box C8	5.65	9.98	3.4E-06	20.1	HOX
Cbfa2t1h	CBFA2T1 identified gene homolog	5.52	9.82	4.5E-05	19.8	other

Foxf2	forkhead box F2	6.38	10.58	1.5E-04	18.3	FH
Peg3	paternally expressed 3	7.05	11.24	9.1E-04	18.3	ZnF
Zfhx1a	Zinc finger homeobox 1a	5.51	9.63	3.2E-06	17.5	HOX
Zfp521	zinc finger protein 521	5.79	9.88	8.3E-06	17.0	ZnF
Myocd	myocardin	6.27	10.35	3.0E-05	17.0	other
Hand2	heart and neural crest derivatives expressed transcript 2	7.18	11.24	9.0E-05	16.6	HLH
Nr2f2	nuclear receptor subfamily 2, group F, member 2	5.71	9.70	4.8E-05	15.9	NHR
Sox11	SRY-box containing gene 11	5.34	9.27	7.2E-08	15.2	HMG
Pbx1	Pre B-cell leukemia transcription factor 1	7.24	11.13	2.6E-03	14.9	HOX
Hoxb3	homeo box B3	7.51	11.27	4.1E-05	13.5	HOX
Bach2	BTB and CNC homology 2	6.53	10.20	5.1E-06	12.7	bZip
Bnc2	basonuclin 2	6.94	10.59	8.3E-05	12.6	ZnF
Satb1	special AT-rich sequence binding protein 1	5.62	9.27	1.1E-04	12.6	HOX
Tshz2	teashirt zinc finger family member 2	7.52	11.08	8.7E-06	11.8	HOX
Nkx2-3	NK2 transcription factor related, locus 3	7.59	11.14	3.5E-05	11.7	HOX
Sox18	SRY-box containing gene 18	8.03	11.50	1.8E-04	11.1	HMG
D14Ertd668e	DNA segment, Chr 14, ERATO Doi 668, expressed	7.27	10.74	3.1E-04	11.0	ZnF
Zfp537	zinc finger protein 537	6.20	9.57	2.1E-06	10.3	HOX
Etv1	ets variant gene 1	5.69	9.06	9.1E-04	10.3	ETS
Tead2	TEA domain family member 2	7.46	10.81	5.7E-04	10.2	other
Ebf1	early B-cell factor 1	5.33	8.67	9.6E-07	10.1	HLH
Hoxb6	homeo box B6	6.64	9.97	3.7E-06	10.0	HOX
Hoxc5	homeo box C5	6.15	9.39	2.1E-05	9.5	HOX
Zfp9	zinc finger protein 9	7.12	10.36	9.9E-04	9.4	ZnF
Erg	Avian erythroblastosis virus E-26 (v-ets) oncogene related	6.77	10.01	7.1E-05	9.4	ETS
Gli3	GLI-Kruppel family member GLI3	6.73	9.94	4.1E-05	9.3	ZnF
Zfhx4	zinc finger homeodomain 4	5.19	8.39	2.7E-04	9.2	ZnF
Nr2f1	nuclear receptor subfamily 2, group F, member 1	5.12	8.31	4.6E-04	9.2	NHR
Tbx2	T-box 2	6.58	9.76	2.9E-04	9.1	other
Hoxd8	homeo box D8	6.15	9.33	2.6E-05	9.1	HOX
Zfp647	zinc finger protein 647	6.41	9.58	2.1E-04	9.0	ZnF

Hivep3	human immunodeficiency virus type I enhancer binding protein 3	5.47	8.65	8.5E-05	9.0	ZnF
Hoxa2	homeo box A2	5.71	8.87	6.2E-04	8.9	HOX
Ascl1	achaete-scute complex homolog-like 1	5.61	8.75	4.0E-05	8.8	HLH
Hlx1	H2.0-like homeo box 1	7.55	10.66	5.9E-05	8.6	HOX
Etv5	ets variant gene 5	6.14	9.21	2.6E-05	8.4	ETS
Vezf1	vascular endothelial zinc finger 1	6.58	9.65	7.5E-04	8.4	ZnF
Hoxb5	homeo box B5	7.91	10.87	3.4E-05	7.8	HOX
Nfib	nuclear factor I/B	7.54	10.49	5.5E-03	7.7	other
Nfia	nuclear factor I/A	8.58	11.51	2.2E-04	7.6	other
Nr1d2	nuclear receptor subfamily 1, group D, member 2	7.02	9.91	1.6E-03	7.4	NHR
Idb4	Inhibitor of DNA binding 4	6.38	9.26	6.3E-04	7.3	HLH
Btbd4	BTB (POZ) domain containing 4	5.46	8.31	6.1E-05	7.2	ZnF
Mef2a	myocyte enhancer factor 2A	6.78	9.62	2.1E-03	7.1	other
Zik1	zinc finger protein interacting with K protein 1	5.39	8.21	1.3E-03	7.0	ZnF
Cbx6	chromobox homolog 6	9.00	11.81	1.3E-05	7.0	other
Zfp62	zinc finger protein 62	6.45	9.25	6.6E-03	7.0	ZnF
Id3	inhibitor of DNA binding 3	8.85	11.63	7.4E-04	6.9	HLH
Nfatc1	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 1	6.92	9.69	9.1E-05	6.8	other
Ebf3	early B-cell factor 3	5.94	8.71	1.5E-03	6.8	HLH
Bhlhb5	basic helix-loop-helix domain containing, class B5	5.59	8.34	3.1E-04	6.7	HLH
Runx2	runt related transcription factor 2	6.35	9.10	3.6E-05	6.7	other
Zfp451	zinc finger protein 451	6.48	9.22	2.8E-04	6.7	ZnF
Phtf2	putative homeodomain transcription factor 2	5.50	8.22	3.4E-04	6.6	ZnF
Hic1	hypermethylated in cancer 1	8.67	11.38	1.9E-04	6.5	ZnF
2810021G02Rik	RIKEN cDNA 2810021G02 gene	6.85	9.54	1.8E-02	6.4	ZnF
Meox2	mesenchyme homeobox 2	5.14	7.83	1.3E-03	6.4	HOX
Notch4	Notch gene homolog 4	7.31	9.98	1.7E-03	6.3	other
Sox7	SRY-box containing gene 7	7.45	10.12	9.2E-05	6.3	HMG
6430601A21Rik	RIKEN cDNA 6430601A21 gene	6.55	9.21	3.3E-03	6.3	ZnF

Zbtb20	zinc finger and BTB domain containing 20	6.77	9.43	8.5E-04	6.3	ZnF
Gli1	GLI-Kruppel family member GLI1	7.78	10.44	1.6E-04	6.3	ZnF
Sox4	SRY-box containing gene 4	9.87	12.52	5.9E-04	6.3	HMG
Notch2	Notch gene homolog 2	7.47	10.11	1.1E-04	6.2	other
A630033E08Rik	RIKEN cDNA A630033E08 gene	6.46	9.08	2.3E-03	6.1	ZnF
Srf	serum response factor	8.50	11.12	3.5E-04	6.1	other
Mxd3	Max dimerization protein 3	6.54	9.13	2.1E-03	6.0	HLH
Zfp37	zinc finger protein 37	6.13	8.71	5.5E-04	6.0	ZnF
Hmg20a	high mobility group 20A	6.59	9.14	5.1E-03	5.9	HMG
Klf12	Kruppel-like factor 12	6.02	8.57	1.1E-03	5.9	ZnF
Hivep2	human immunodeficiency virus type I enhancer binding protein 2	6.93	9.47	2.6E-03	5.8	ZnF
Ahr	aryl-hydrocarbon receptor	8.95	11.47	1.8E-03	5.7	HLH
Zfp260	zinc finger protein 260	8.54	11.06	1.3E-03	5.7	ZnF
Zfp161	zinc finger protein 161	7.46	9.96	1.2E-02	5.7	ZnF
Asb4	ankyrin repeat and SOCS box-containing protein 4	6.35	8.81	2.3E-04	5.5	other
Klf2	Kruppel-like factor 2	7.82	10.27	2.1E-04	5.5	ZnF
Atbf1	AT motif binding factor 1	7.10	9.54	1.1E-03	5.4	HOX
Aprin	androgen-induced proliferation inhibitor	8.04	10.47	1.8E-03	5.4	other
Zbtb4	zinc finger and BTB domain containing 4	6.68	9.09	2.6E-03	5.3	ZnF
Smarcd3	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 3	7.90	10.31	5.1E-04	5.3	other
2610305D13Rik	RIKEN cDNA 2610305D13 gene	5.30	7.70	2.4E-03	5.3	ZnF
Zfp532	zinc finger protein 532	8.48	10.87	1.4E-05	5.2	ZnF
Hoxa4	homeo box A4	7.27	9.63	2.0E-04	5.1	HOX
Zfp423	zinc finger protein 423	6.60	8.94	1.9E-05	5.0	ZnF
Phox2a	paired-like homeobox 2a	6.91	9.23	8.7E-04	5.0	HOX
Glis2	GLIS family zinc finger 2	8.58	10.89	1.4E-04	5.0	ZnF
Mecp2	methyl CpG binding protein 2	7.31	9.61	5.4E-03	4.9	other
Plagl1	pleiomorphic adenoma gene-like 1	8.94	11.24	4.2E-03	4.9	ZnF
Zfp287	zinc finger protein 287	7.64	9.93	1.0E-03	4.9	ZnF

Notch3	Notch gene homolog 3	8.95	11.22	1.8E-03	4.8	other
G630024C07Rik	RIKEN cDNA G630024C07 gene	6.94	9.19	1.0E-03	4.8	ZnF
Tcf3	transcription factor 3	7.82	10.07	9.6E-04	4.7	HMG
Sdccag33l	Serologically defined colon cancer antigen 33 like	7.15	9.40	4.5E-05	4.7	HOX
Zfp536	zinc finger protein 536	6.51	8.75	4.3E-04	4.7	ZnF
Hoxd3	homeo box D3	7.54	9.77	6.6E-04	4.7	HOX
MGI:1932093	Jun dimerization protein 2	7.85	10.08	1.7E-03	4.7	bZip
Zfp367	zinc finger protein 367	7.70	9.92	3.1E-03	4.7	ZnF
Foxc2	forkhead box C2	5.89	8.09	9.2E-04	4.6	FH
Hhex	hematopoietically expressed homeobox	8.11	10.31	1.5E-04	4.6	HOX
Zbed3	zinc finger, BED domain containing 3	6.27	8.45	6.6E-03	4.5	ZnF
Zfpn1a2	zinc finger protein, subfamily 1A, 2	6.05	8.22	2.0E-03	4.5	ZnF
Sall2	sal-like 2	6.50	8.65	4.5E-04	4.4	ZnF
Zfp583	zinc finger protein 583	5.84	7.98	1.7E-04	4.4	ZnF
Zfp644	zinc finger protein 644	7.73	9.87	1.1E-02	4.4	ZnF
Rora	RAR-related orphan receptor alpha	5.89	8.03	5.1E-03	4.4	NHR
Nfatc4	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 4	6.21	8.33	3.6E-04	4.4	other
Zfp449	zinc finger protein 449	7.38	9.47	6.7E-04	4.3	ZnF
Hoxb4	homeo box B4	7.36	9.45	1.1E-04	4.3	HOX
Zfpm2	zinc finger protein, multitype 2	6.68	8.77	3.6E-03	4.3	ZnF
Zfp462	zinc finger protein 462	7.61	9.70	4.3E-03	4.3	ZnF
Sox2	SRY-box containing gene 2	5.22	7.30	1.9E-03	4.2	HMG
Tcf15	transcription factor 15	7.57	9.64	9.9E-04	4.2	HLH
Zfp184	zinc finger protein 184 (Kruppel-like)	6.75	8.81	6.1E-05	4.2	ZnF
Tox	thymocyte selection-associated HMG box gene	6.48	8.54	3.6E-02	4.2	HMG
Ilf2	interleukin enhancer binding factor 2	7.10	9.15	1.4E-02	4.1	other
Zfp105	zinc finger protein 105	7.08	9.13	8.2E-06	4.1	ZnF
Glis3	GLIS family zinc finger 3	5.58	7.63	1.1E-03	4.1	ZnF
Pcgf4	polycomb group ring finger 4	7.84	9.88	2.4E-02	4.1	ZnF
Foxj2	forkhead box J2	8.60	10.62	6.3E-03	4.1	FH

Hoxb2	homeo box B2	8.55	10.57	1.2E-04	4.1	HOX
Hoxb7	homeo box B7	6.40	8.42	1.9E-03	4.0	HOX
Zfp286	zinc finger protein 286	5.93	7.93	8.5E-05	4.0	ZnF
Wig1	wild-type p53-induced gene 1	8.53	10.52	2.9E-03	4.0	ZnF
Scml2	sex comb on midleg-like 2	6.47	8.45	7.3E-04	3.9	other
Bcl11a	B-cell CLL/lymphoma 11A (zinc finger protein)	6.09	8.07	7.2E-03	3.9	ZnF
Tal1	T-cell acute lymphocytic leukemia 1	6.99	8.95	9.2E-04	3.9	HLH
Msc	musculin	8.29	10.25	9.5E-05	3.9	HLH
Phf6	PHD finger protein 6	7.80	9.76	2.1E-03	3.9	ZnF
Zbtb16	zinc finger and BTB domain containing 16	5.59	7.53	5.3E-03	3.9	ZnF
3100002L24Rik	RIKEN cDNA 3100002L24 gene	7.68	9.63	5.0E-03	3.9	ZnF
Pias1	protein inhibitor of activated STAT 1	6.64	8.57	3.2E-02	3.8	other
2810426N06Rik	RIKEN cDNA 2810426N06 gene	6.34	8.25	2.3E-02	3.8	ZnF
Hmgn1	high mobility group nucleosomal binding domain 1	11.73	13.61	1.8E-03	3.7	HMG
Zfp68	zinc finger protein 68	8.51	10.38	8.5E-03	3.6	ZnF
Zfp278	zinc finger protein 278	8.39	10.22	1.4E-03	3.6	ZnF
BC029103	cDNA sequence BC029103	7.86	9.69	9.1E-03	3.6	ZnF
Tsc22d1	TSC22 domain family, member 1	11.17	12.99	7.2E-04	3.5	other
Zfp354c	zinc finger protein 354C	7.86	9.67	4.6E-03	3.5	ZnF
Zfp46	zinc finger protein 46	10.02	11.82	6.4E-04	3.5	ZnF
Hoxc4	homeo box C4	8.04	9.84	1.4E-02	3.5	HOX
Laf4	AF4/FMR2 family, member 3 (Aff3)	8.19	10.00	3.5E-04	3.5	bZip
BC043476	cDNA sequence BC043476	6.01	7.81	6.7E-03	3.5	ZnF
Zfp61	zinc finger protein 61	6.94	8.74	1.2E-03	3.5	ZnF
3100002L24Rik	RIKEN cDNA 3100002L24 gene	9.20	11.00	4.7E-03	3.5	ZnF
Hmgb1	high mobility group box 1	9.84	11.64	8.6E-03	3.5	HMG
Tead1	TEA domain family member 1	5.37	7.16	4.9E-02	3.4	other
Gli2	GLI-Kruppel family member GLI2	9.39	11.17	1.3E-04	3.4	ZnF
Whsc1	Wolf-Hirschhorn syndrome candidate 1	7.64	9.43	1.2E-02	3.4	other
Zbtb1	Zinc finger and BTB domain containing 1	6.08	7.85	1.0E-02	3.4	ZnF
Mta1	metastasis associated 1	8.87	10.64	5.0E-03	3.4	ZnF

Baz1b	bromodomain adjacent to zinc finger domain, 1B	8.04	9.81	1.3E-02	3.4	ZnF
Phtf1	putative homeodomain transcription factor 1	8.26	10.02	6.3E-03	3.4	HOX
Setbp1	SET binding protein 1	9.18	10.93	1.8E-03	3.4	other
Zfp90	zinc finger protein 90	7.01	8.75	2.1E-03	3.4	ZnF
Cutl1	Cut-like 1, transcript variant 2	8.32	10.07	2.2E-03	3.4	HOX
Foxm1	forkhead box M1	8.11	9.85	1.1E-03	3.3	FH
Foxk2	forkhead box K2	8.84	10.57	2.4E-03	3.3	FH
Zfp60	zinc finger protein 60	6.23	7.96	2.6E-02	3.3	ZnF
Tbx1	T-box 1	6.49	8.21	4.5E-03	3.3	other
Zfp606	zinc finger protein 606	5.92	7.64	7.9E-03	3.3	ZnF
5730601F06Rik	RIKEN cDNA 5730601F06 gene	7.84	9.56	2.9E-02	3.3	ZnF
Npas4	neuronal PAS domain protein 4	6.27	7.98	1.6E-04	3.3	other
Gata2	GATA binding protein 2	6.91	8.62	5.4E-04	3.3	ZnF
Zfp326	zinc finger protein 326	8.50	10.21	4.8E-03	3.3	ZnF
Zfp386	zinc finger protein 386 (Kruppel-like)	7.80	9.50	9.1E-03	3.3	ZnF
Phf20l1	PHD finger protein 20-like 1	7.59	9.28	2.0E-02	3.2	ZnF
Ahctf1	AT hook containing transciption factor 1	7.60	9.29	1.9E-02	3.2	other
2610044O15Rik	RIKEN cDNA 2610044O15 gene	5.81	7.50	1.3E-02	3.2	ZnF
Yaf2	YY1 associated factor 2	9.11	10.79	9.6E-03	3.2	ZnF
Mll	myeloid/lymphoid or mixed-lineage leukemia	7.71	9.39	2.6E-03	3.2	ZnF
Nr4a2	Nuclear receptor subfamily 4, group A, member 2	5.71	7.37	3.4E-03	3.2	NHR
Tcf12	Transcription factor 12	7.37	9.02	2.5E-02	3.2	HLH
Sox12	SRY-box containing gene 12	9.41	11.06	9.9E-04	3.1	HMG
Pbx2	pre B-cell leukemia transcription factor 2	8.54	10.19	2.3E-03	3.1	HOX
Zfp275	Zinc finger protein 275	9.32	10.96	7.3E-04	3.1	ZnF
Atf2	activating transcription factor 2	9.35	10.99	2.5E-02	3.1	ZnF
Notch1	Notch gene homolog 1	9.45	11.09	1.8E-03	3.1	other
Rarg	retinoic acid receptor, gamma	9.25	10.88	5.5E-03	3.1	NHR
BC052046	CDNA sequence BC052046	7.15	8.78	2.0E-03	3.1	ZnF
9130211I03Rik	RIKEN cDNA 9130211I03 gene	8.13	9.74	5.0E-04	3.1	bZip
Bcl6b	B-cell CLL/lymphoma 6, member B	8.18	9.80	2.0E-04	3.1	ZnF

Zfp74	zinc finger protein 74	6.76	8.37	2.1E-03	3.1	ZnF
Zfpn1a5	zinc finger protein, subfamily 1A, 5	7.98	9.59	2.6E-02	3.0	ZnF
Zfp592	zinc finger protein 592	7.89	9.50	1.6E-02	3.0	ZnF
6030408C04Rik	RIKEN cDNA 6030408C04 gene	8.28	9.88	2.6E-02	3.0	ZnF
Lhx6	LIM homeobox protein 6	6.89	8.47	1.2E-03	3.0	HOX
C630016O21Rik	RIKEN cDNA C630016O21 gene	6.84	8.42	1.7E-02	3.0	ZnF
BC031407	GATA zinc finger domain containing 2A	7.54	9.11	1.2E-02	3.0	ZnF
Hey1	hairy/enhancer-of-split related with YRPW motif 1	8.82	10.39	6.8E-04	3.0	HLH
Lyl1	lymphoblastomic leukemia	8.12	9.69	1.9E-03	3.0	HLH
Zfp36l1	zinc finger protein 36, C3H type-like 1	8.97	10.52	6.0E-04	2.9	ZnF
Nfat5	Nuclear factor of activated T-cells 5	6.27	7.82	3.1E-02	2.9	other
Rybp	RING1 and YY1 binding protein	9.97	11.51	6.0E-03	2.9	ZnF
Hoxa3	homeo box A3	7.67	9.22	8.0E-04	2.9	HOX
Jmjd2c	jumonji domain containing 2C	7.06	8.60	1.1E-03	2.9	other
D130064H19Rik	RIKEN cDNA D130064H19 gene	8.66	10.19	2.6E-02	2.9	other
Epas1	endothelial PAS domain protein 1	8.65	10.18	1.9E-03	2.9	HLH
Lmo4	LIM domain only 4	7.19	8.71	1.4E-02	2.9	other
Lmo2	LIM domain only 2	9.59	11.11	1.3E-03	2.9	other
Zfp202	zinc finger protein 202	7.67	9.20	6.8E-03	2.9	ZnF
Ctcf	CCCTC-binding factor	6.44	7.96	4.6E-03	2.9	HOX
Zipro1	zinc finger proliferation 1	6.68	8.19	1.7E-02	2.8	ZnF
Sdccag33	serologically defined colon cancer antigen 33	9.52	11.02	1.2E-02	2.8	ZnF
BC043301	cDNA sequence BC043301	5.76	7.26	1.1E-03	2.8	ZnF
Ilf3	interleukin enhancer binding factor 3	10.09	11.59	1.1E-02	2.8	ZnF
Nfic	nuclear factor I/C	6.33	7.82	1.7E-02	2.8	other
MGI:1927369	RB-associated KRAB repressor	7.92	9.40	7.8E-03	2.8	ZnF
Rarb	retinoic acid receptor, beta	7.68	9.17	2.1E-02	2.8	NHR
Scmh1	Sex comb on midleg homolog 1	7.35	8.83	2.0E-02	2.8	other
Bnc1	basonuclin 1	6.75	8.23	4.4E-04	2.8	ZnF
Cbfb	core binding factor beta	10.27	11.75	7.1E-03	2.8	other
Egr1	early growth response 1	8.73	10.21	8.1E-04	2.8	ZnF

BC002059	CDNA sequence BC002059	6.38	7.84	5.3E-03	2.8	ZnF
MGI:3028594	zinc finger protein 422, related sequence 1	8.76	10.22	1.8E-03	2.8	ZnF
Btbd11	BTB (POZ) domain containing 11	6.30	7.76	2.5E-03	2.7	other
1110005A23Rik	RIKEN cDNA 1110005A23 gene	8.46	9.92	2.4E-02	2.7	other
Hmgn2	high mobility group nucleosomal binding domain 2	12.33	13.78	2.7E-03	2.7	HMG
Prox1	prospero-related homeobox 1	5.66	7.11	1.2E-02	2.7	HOX
Hmgb1	high mobility group box 1	12.29	13.74	3.7E-03	2.7	HMG
Phf20	PHD finger protein 20	8.50	9.94	4.3E-03	2.7	ZnF
Zfpn1a1	zinc finger protein, subfamily 1A, 1	7.46	8.90	2.1E-03	2.7	ZnF
Hoxc6	homeo box C6	7.47	8.91	5.2E-03	2.7	HOX
Phf14	PHD finger protein 14	6.33	7.77	1.8E-02	2.7	ZnF
Zfp82	zinc finger protein 82	6.63	8.06	1.5E-02	2.7	ZnF
D930038J03Rik	RIKEN cDNA D930038J03 gene	9.64	11.07	1.3E-02	2.7	ZnF
Crem	cAMP responsive element modulator	6.98	8.40	1.4E-02	2.7	bZip
Zbtb41	zinc finger and BTB domain containing 41 homolog	8.40	9.82	9.1E-03	2.7	ZnF
Bard1	BRCA1 associated RING domain 1	6.53	7.95	3.3E-04	2.7	ZnF
Zfp397	zinc finger protein 397	7.31	8.72	2.0E-02	2.7	ZnF
Tcf7	transcription factor 7, T-cell specific	7.38	8.79	1.1E-02	2.7	HMG
BC050078	cDNA sequence BC050078	5.70	7.10	1.1E-04	2.7	ZnF
Rbpsuh	recombining binding protein suppressor of hairless	8.60	10.00	4.3E-03	2.6	other
Runx1	runt related transcription factor 1	7.01	8.40	4.6E-03	2.6	other
9530033F24Rik	RIKEN cDNA 9530033F24 gene	7.02	8.41	3.7E-03	2.6	ZnF
Sp2	Sp2 transcription factor	7.82	9.20	2.3E-02	2.6	ZnF
Zmynd11	zinc finger, MYND domain containing 11	9.27	10.65	4.8E-03	2.6	ZnF
5830435C13Rik	RIKEN cDNA 5830435C13 gene	8.91	10.28	4.2E-02	2.6	ZnF
Bhc80	PHD finger protein 21A (Phf21a)	9.71	11.07	2.2E-03	2.6	ZnF
Fem1c	fem-1 homolog c	9.16	10.52	2.2E-02	2.6	other
Zfp637	zinc finger protein 637	8.89	10.26	5.1E-04	2.6	ZnF
Hivep1	human immunodeficiency virus type I enhancer binding protein 1	8.27	9.63	2.0E-02	2.6	ZnF
Zfp40	zinc finger protein 40	5.72	7.08	1.1E-02	2.6	ZnF

Pknox1	Pbx/knotted 1 homeobox	8.26	9.62	1.4E-02	2.6	HOX
Smarce1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1	9.34	10.69	1.2E-03	2.6	HMG
Ebf4	Early B-cell factor 4	7.97	9.32	4.8E-03	2.5	HLH
Meox1	mesenchyme homeobox 1	9.19	10.54	2.1E-03	2.5	HOX
Zfp131	zinc finger protein 131	8.88	10.23	2.8E-02	2.5	ZnF
Foxk1	forkhead box K1	8.34	9.68	1.0E-02	2.5	FH
Hif1a	Hypoxia inducible factor 1, alpha subunit	6.10	7.44	2.6E-02	2.5	HLH
Hmx3	H6 homeo box 3	6.02	7.36	3.8E-03	2.5	HOX
Hey2	hairy/enhancer-of-split related with YRPW motif 2	5.34	6.66	1.9E-03	2.5	HLH
Rbl2	retinoblastoma-like 2	7.83	9.16	3.7E-02	2.5	other
Pogk	pogo transposable element with KRAB domain	8.48	9.81	1.1E-03	2.5	ZnF
Zfp192	Zinc finger protein 192	8.70	10.03	2.2E-02	2.5	ZnF
Gatad2b	GATA zinc finger domain containing 2B	9.13	10.45	2.4E-03	2.5	ZnF
Brd8	bromodomain containing 8	7.53	8.85	1.6E-02	2.5	other
Hlf	hepatic leukemia factor	6.77	8.09	2.7E-04	2.5	bZip
Zfp322a	zinc finger protein 322a	9.46	10.77	2.4E-03	2.5	ZnF
Mxd4	Max dimerization protein 4	8.56	9.87	9.7E-04	2.5	HLH
E2f7	E2F transcription factor 7	8.57	9.87	2.8E-02	2.5	other
Zfp516	zinc finger protein 516	8.97	10.27	1.8E-02	2.5	ZnF
Lass4	longevity assurance homolog 4	8.10	9.40	1.4E-02	2.5	HOX
Zfp639	zinc finger protein 639	8.87	10.16	1.0E-02	2.5	ZnF
Zfp142	zinc finger protein 142	6.38	7.68	2.0E-03	2.5	ZnF
Creb3l2	cAMP responsive element binding protein 3-like 2	9.53	10.81	1.4E-02	2.4	bZip
Tcf19	transcription factor 19	9.65	10.93	7.4E-04	2.4	other
Zfp30	zinc finger protein 30	8.14	9.41	2.2E-04	2.4	ZnF
Hmgb2	high mobility group box 2	11.72	12.99	6.5E-03	2.4	HMG
Cited4	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 4	6.23	7.49	1.3E-02	2.4	other
Tgif2	TGFB-induced factor 2	8.09	9.36	1.6E-03	2.4	HOX
Foxp1	Forkhead box P1	10.05	11.31	3.1E-03	2.4	FH

1110051B16Rik	RIKEN cDNA 1110051B16 gene	7.38	8.63	4.0E-02	2.4	HOX
Arid5b	Modulator recognition factor 2 (Mrf2)	10.12	11.37	5.3E-03	2.4	other
Zfp191	zinc finger protein 191	9.32	10.55	5.2E-02	2.4	ZnF
Klf9	Kruppel-like factor 9	5.67	6.91	2.3E-02	2.4	ZnF
9830124H08Rik	RIKEN cDNA 9830124H08 gene	8.59	9.82	1.4E-02	2.3	ZnF
Myt1l	myelin transcription factor 1-like	5.42	6.64	8.1E-05	2.3	ZnF
Dmtf1	cyclin D binding myb-like transcription factor 1	9.28	10.50	4.0E-02	2.3	other
Sox5	SRY-box containing gene 5	6.99	8.21	2.7E-03	2.3	HMG
Id2	inhibitor of DNA binding 2	7.10	8.31	1.4E-02	2.3	HLH
Phf2	PHD finger protein 2	9.20	10.41	5.4E-03	2.3	ZnF
Nfyb	nuclear transcription factor-Y beta	7.82	9.03	1.1E-02	2.3	other
1110034O07Rik	RIKEN cDNA 1110034O07 gene	9.17	10.37	9.7E-03	2.3	ZnF
Arid2	AT rich interactive domain 2 (Arid-rfx like)	8.49	9.70	3.9E-03	2.3	ZnF
Twist2	twist homolog 2	8.77	9.98	5.1E-04	2.3	HLH
Arnt2	aryl hydrocarbon receptor nuclear translocator 2	7.70	8.90	3.2E-02	2.3	HLH
6330581L23Rik	RIKEN cDNA 6330581L23 gene	6.58	7.78	2.9E-02	2.3	ZnF
Pknox2	Pbx/knotted 1 homeobox 2	7.91	9.11	2.8E-04	2.3	HOX
Rlf	rearranged L-myc fusion sequence	8.34	9.53	2.7E-02	2.3	ZnF
Myt1	myelin transcription factor 1	6.79	7.98	3.0E-03	2.3	ZnF
Rfx3	Regulatory factor X, 3 (influences HLA class II expression)	6.30	7.48	3.0E-02	2.3	HLH
Fhl2	four and a half LIM domains 2	10.21	11.39	7.8E-03	2.3	other
Zbtb34	zinc finger and BTB domain containing 34	8.75	9.93	2.0E-03	2.3	ZnF
Dlx5	distal-less homeobox 5	6.91	8.09	6.8E-03	2.3	HOX
Smad1	MAD homolog 1	9.67	10.84	1.0E-02	2.3	other
2310030G09Rik	RIKEN cDNA 2310030G09 gene	5.71	6.88	2.3E-03	2.3	ZnF
Zfp454	zinc finger protein 454	6.70	7.87	3.0E-02	2.2	ZnF
Insm1	insulinoma-associated 1	6.22	7.39	2.9E-02	2.2	ZnF
Jmjd2a	jumonji domain containing 2A	6.26	7.43	4.6E-03	2.2	other
Creb5	CAMP responsive element binding protein 5	6.85	8.01	2.5E-03	2.2	bZip
Ncoa3	Nuclear receptor coactivator 3	6.86	8.02	1.1E-02	2.2	HLH
Tieg3	TGFB inducible early growth response 3	8.74	9.90	7.3E-03	2.2	ZnF

Jarid1a	jumonji, AT rich interactive domain 1A (Rbp2 like)	8.43	9.58	5.4E-02	2.2	other
Cxxc1	CXXC finger 1 (PHD domain)	7.74	8.89	2.4E-02	2.2	ZnF
Ash1l	ash1 (absent, small, or homeotic)-like	9.15	10.29	1.3E-02	2.2	ZnF
Klf7	Kruppel-like factor 7 (ubiquitous)	8.91	10.06	5.4E-03	2.2	ZnF
Elf2	E74-like factor 2	9.61	10.75	1.6E-02	2.2	ETS
Trim28	tripartite motif protein 28	10.98	12.12	8.9E-03	2.2	ZnF
Zfp26	zinc finger protein 26	9.48	10.62	3.5E-02	2.2	ZnF
Zzz3	zinc finger, ZZ domain containing 3	9.57	10.71	3.2E-03	2.2	other
Zfp329	zinc finger protein 329	7.73	8.85	2.2E-02	2.2	ZnF
Zfp354a	zinc finger protein 354A	6.65	7.76	3.8E-02	2.2	ZnF
Zfp472	zinc finger protein 472	7.45	8.56	2.8E-02	2.2	ZnF
Mga	MAX gene associated	9.21	10.33	1.3E-02	2.2	HLH
Smad6	MAD homolog 6	9.54	10.66	3.0E-02	2.2	other
Zfp334	zinc finger protein 334	8.68	9.79	2.1E-02	2.1	ZnF
Ncoa2	nuclear receptor coactivator 2	7.81	8.91	2.5E-03	2.1	HLH
3000002G13Rik	RIKEN cDNA 3000002G13 gene	7.02	8.12	5.5E-02	2.1	ZnF
Zfp553	zinc finger protein 553	7.56	8.66	2.7E-02	2.1	ZnF
Six5	sine oculis-related homeobox 5 homolog	9.12	10.21	4.2E-06	2.1	HOX
Mynn	myoneurin	8.66	9.75	6.0E-03	2.1	ZnF
Wt1	Wilms tumor homolog	7.00	8.09	1.7E-02	2.1	ZnF
Zfp87	zinc finger protein 87	7.93	9.01	2.2E-02	2.1	ZnF
Gmeb1	glucocorticoid modulatory element binding protein 1	7.75	8.83	4.8E-02	2.1	other
Gatad2a	GATA zinc finger domain containing 2A	8.30	9.38	9.4E-03	2.1	ZnF
Zfp236	zinc finger protein 236	7.77	8.84	2.7E-02	2.1	ZnF
Dlx2	distal-less homeobox 2	6.28	7.35	1.7E-02	2.1	HOX
E2f1	E2F transcription factor 1	7.79	8.86	1.4E-02	2.1	other
Nrf1	Nuclear respiratory factor 1	8.30	9.37	8.7E-03	2.1	other
Hes5	hairy and enhancer of split 5	5.92	6.99	4.0E-04	2.1	HLH
Tead3	TEA domain family member 3	9.20	10.26	2.8E-03	2.1	other
Zfp148	zinc finger protein 148	8.35	9.41	2.5E-02	2.1	ZnF
Phf13	PHD finger protein 13	8.67	9.72	1.9E-02	2.1	ZnF

Zfp422	zinc finger protein 422	10.44	11.49	1.3E-02	2.1	ZnF
Zfp41	zinc finger protein 41	8.98	10.03	1.3E-03	2.1	ZnF
Zfp59	zinc finger protein 59	7.75	8.79	9.1E-03	2.1	ZnF
St18	suppression of tumorigenicity 18	5.54	6.59	1.4E-03	2.1	ZnF
Mll5	myeloid/lymphoid or mixed-lineage leukemia 5	8.41	9.45	5.3E-02	2.1	ZnF
Zfp282	zinc finger protein 282	6.60	7.64	1.5E-02	2.0	ZnF
Nfya	nuclear transcription factor-Y alpha	8.53	9.56	1.3E-02	2.0	other
Smad7	MAD homolog 7	9.67	10.70	1.8E-02	2.0	other
Ets2	E26 avian leukemia oncogene 2, 3' domain	10.98	12.01	1.3E-02	2.0	ETS
Crebl1	cAMP responsive element binding protein-like 1	8.32	9.33	2.6E-02	2.0	bZip
Tcf7l2	transcription factor 7-like 2, T-cell specific, HMG-box	7.31	8.32	2.6E-02	2.0	HMG
Ches1	checkpoint supressor 1	9.14	10.14	3.4E-02	2.0	FH
Nr2c1	nuclear receptor subfamily 2, group C, member 1	7.62	8.63	4.6E-02	2.0	NHR
Id1	inhibitor of DNA binding 1	10.67	11.68	4.7E-03	2.0	HLH
Tcfap2a	transcription factor AP-2, alpha	8.18	9.18	1.8E-02	2.0	other

[a]Epi = Average expression value in epithelium (Log2) as calculated by RMA.
[b]Mes = Average expression value in mesenchyme (Log2) as calculated by RMA.
[c]ES = Enrichment score (absolute value of numerical fold difference, calculated as in Materials and Methods).
[d]TF-Family = Transcription factor family.

Table 2.7 Epithelial genes with Hnf4 binding sites.

Gene Symbol	TSS[a]	Start[b]	End[c]	Core Sim[d]	Matrix Sim[e]	Sequence
Otc	1	371	391	1	0.78	ttaggcttAAAGttcaagtgc
Otc	1	451	471	1	0.90	gagggagCAAAggtcttagca
Slc26a3	1	276	296	1	0.84	agagactCAAAggtcaagacc
Gjb1	1	20	40	1	0.85	ctgaggtCAAAgtgggagatg
Gjb1	1	424	444	1	0.83	gtactgtCAAAgcccaaccca
2010106E10Rik	1	433	453	0.75	0.85	gtaagggAAAAgttcaaaatc
Leap2	1	407	427	1	0.89	gatggggCAAAgtttgtttgt
Rnf128	1	110	130	1	0.83	ccaggtcCAAAgggtgtgagt
BC013481	1	157	177	0.75	0.79	gccgctcaAAAAgtcacatcc
BC013481	1	171	191	1	0.93	gagcggcCAAAgttcgtaaat
BC013481	1	280	300	0.75	0.84	cgctggcCGAAggccagtgca
BC013481	2	73	93	1	0.83	ccaagatCAAAgcacctggag
BC013481	2	196	216	1	0.86	ccctggaCAAAggaagcatgc
BC013481	2	358	378	1	0.93	tgtgggtCAAAgttccatcaa
BC013481	2	412	432	0.75	0.84	ccctgggCAATggccaggcta
Ace2	1	436	456	1	0.83	ccaagttCAAAggctgatgag
Serpina1a	1	280	300	1	0.86	ctgtgtcCAAAgtgcgtctgg
Serpina1a	1	379	399	0.75	0.84	ttcgggaCAGAgctcagctct
Serpina1a	1	426	446	0.75	0.84	agaggggCTAAgtccatcgag
Serpina1a	2	315	335	0.75	0.84	ttcgggaCAGAgctcagctct
Serpina1a	2	362	382	0.75	0.86	agaggggCTAAgttcatcgat
Ugt2b5	1	447	467	0.75	0.82	cttggatAAAAggtcaagcag
Afp	1	325	345	1	0.94	tcaaggaCAAAgaccacttca
Cyp3a13	1	273	293	1	0.88	ctgagtaCAAAggttattgct
Olfm4	1	144	164	1	0.93	agaaggaCAAAggttactatt
Olfm4	2	121	141	1	0.93	agaaggaCAAAggttactatt
Reep6/Dp1l1	1	305	325	1	0.86	gggcggaCAAAggaggtgggg
Reep6/Dp1l1	1	373	393	1	0.77	cggggccgAAAGcactgggcc
Abcc2	1	346	366	1	0.83	tcaggatgAAAGgtcaggaga
Pck1	1	46	66	1	0.98	ccacggcCAAAggtcatgaga
Pck1	1	228	248	1	0.92	gcagggtCAAAgtttagtcaa
Defcr12	1	330	350	1	0.85	ttgaggaCAAAggaacatcaa
Tm6sf2	1	436	456	1	0.90	ccgcgcgCAAAggtcgccccg
Clca6	1	411	431	0.75	0.85	taaaggtCAATgttcatgtac
Clca3	1	350	370	0.75	0.79	tagggtcaAGAGgttaaaggc
Slc7a9	1	76	96	1	0.77	ctggttccAAAGtttattgct
Slc7a9	1	346	366	0.75	0.82	gaaggtaCAGAggtcaacaga
Mep1b	1	406	426	1	0.90	ttaagatCAAAggccggaagt
Mep1a	1	465	485	1	0.88	gcacggtCAAAgtttgcaccg
Slc3a1	1	426	446	1	0.94	ggctggaCAAAgtccagccat
Mgam	1	421	441	1	0.85	gtaatgcCAAAgctcagctgt

Hsd17b13	1	348	368	1	0.92	ccaggtcCAAAgggcacatcc
Muc13	1	217	237	1	0.93	gcaaggtCAAAggtgagtgta
Hnf4g	1	461	481	1	0.83	cattgagCAAAgctaatattc
Ugt2b34	1	366	386	1	0.85	tggagggCAAAtgccaaaacc
Anxa13	1	397	417	1	0.86	gtgtggaCAAAggtgtgatct
Cubn	1	272	292	0.75	0.76	aggggtcaAAATctaccttaa
Cubn	1	396	416	1	0.87	tgctgatCAAAgagcaactgg
Ehhadh	1	54	74	1	0.84	tggttgcCAAAgtccaccaag
Perp	1	77	97	1	0.87	caaggtccAAAGgtagatccg
Mfi2	1	397	417	0.75	0.78	gccggtccAAGGgttgctgac
Gda	1	50	70	1	0.81	gatgggcaAAAGgactaaaat
Gda	1	239	259	0.75	0.87	gacaggtGAAAggtcactttg
BC040758	1	277	297	0.75	0.86	agaggacaGAAGgtcctgaca
Ckmt1	1	293	313	0.75	0.86	caagggcaTAAGgtcatcagg
Chpt1	1	131	151	1	0.82	acctgagCAAAggctggagag
Tspan8	1	274	294	1	0.86	aggagcaCAAAgttcctttca
Tspan8	2	261	281	1	0.86	aggagcaCAAAgttcctttca
Ocln	1	97	117	0.75	0.82	agagggtCCAAgggcgcctgg
Ocln	1	243	263	0.75	0.81	aagggtctCAAGatcgtggag
Ocln	2	134	154	0.75	0.81	aagggtctCAAGatcgtggag
Slc2a2	1	291	311	1	0.88	acagagaCAAAgttcaaagca
Hmgcs2	1	391	411	1	0.94	actgggcCAAAggtctcagaa
Ces5	1	342	362	1	0.93	aggaggcCAAAgagcaggaat
Ces5	1	403	423	1	0.95	ggcggggCAAAgttcgtattg
2610528J11Rik	1	197	217	1	0.85	caaagttCAAAgttgaagccg
2610528J11Rik	1	204	224	1	0.93	gccggtaCAAAgttcaaagtt
Guca2b	1	477	497	0.75	0.81	cagggtccAGAGctctggtca
Vil1	1	67	87	0.75	0.85	ccaaggcCAGAgttcacactt
Vil1	1	443	463	1	0.97	gtgaggaCAAAggtcgttcgg
Gpr151	1	408	428	0.75	0.77	gagggtccAGAGgtgtcttca
Spink3	1	234	254	1	0.85	gggaagtCAAAggtctccttt
Spink3	1	340	360	1	0.83	ttgggtccAAAGttttctgcc
Myo5b	1	91	111	1	0.81	ggtgggcaAAAGtgccatggt
Myo5b	1	437	457	1	0.80	cagggagtAAAGgtcgcgcgg
Myo5b	2	32	52	1	0.82	acaggcaCAAAgagctgtaac
Tacstd1	1	220	240	1	0.83	cgcagcgCAAAgtcaagtatt
Apob	1	397	417	1	0.92	aaaggtcCAAAgggcacggcc
Apob	1	405	425	1	0.82	gaattgcCAAAggtccaaagg
Apob	2	291	311	1	0.92	aaaggtcCAAAgggcacggcc
Apob	2	299	319	1	0.82	gaattgcCAAAggtccaaagg
Apob	3	358	378	1	0.84	tgtagtaCAAAgtcaagcaca
Anxa4	1	148	168	0.75	0.79	atgggccaGAAGctcacaaga
4732466D17Rik	1	150	170	0.75	0.83	caaggacaCAAGgtcacatct
4732466D17Rik	1	314	334	1	0.82	atgagagCAAAggtgtgtgtg

[a]TSS = Transcription Start Site. Numbers indicate that more than one promoter were identified for the gene. Each site is assigned to a specific TSS (number given) in a specific promoter.

[b]Start = Beginning of the Hnf4-binding site relative to the TSS (all sites are upstream of the TSS).

[c]End = Ending of the Hnf4-binding site relative to the TSS.

[d]CoreSim = Core similarity. Initial MatInspector search is done with an optional preselection in which only matches to the core region are considered. This reduces the total number of matches and simultaneously accelerates the performance of the search.

[e]MatrixSim = Matrix similarity

The matrix similarity is calculated only if the core similarity reaches a user defined threshold (core similarity). Further definition of the Core and Matrix Similarities are given in the online help page of MatInspector: http://www.genomatix.de/online_help/help_matinspector/matinspector_alg.html

Table 2.8 Compartmentalization of signaling pathway genes.

Symbol	Gene Title	Epi[a]	Mes[b]	P-value	ES[c]	E/M[d]
Notch Pathway						
Dll4	delta-like 4	7.99	8.99	1.1E-03	2.0	Mes
Dner	delta/notch-like EGF-related receptor	5.28	8.01	8.0E-04	6.6	Mes
Jag1	jagged 1	7.69	10.58	1.8E-04	7.4	Mes
Jag2	jagged 2	7.85	8.86	2.7E-02	2.0	Mes
Notch1	Notch gene homolog 1	9.45	11.09	1.8E-03	3.1	Mes
Notch2	Notch gene homolog 2	7.47	10.11	1.1E-04	6.2	Mes
Notch3	Notch gene homolog 3	8.95	11.22	1.8E-03	4.8	Mes
Notch4	Notch gene homolog 4	7.31	9.98	1.7E-03	6.3	Mes
Hedgehog Pathway						
Ihh	Indian hedgehog	9.99	8.52	1.5E-03	2.8	Epi
Ptch1	patched homolog 1	8.28	13.08	2.8E-04	27.7	Mes
Ptch2	patched homolog 2	7.60	9.34	2.1E-03	3.4	Mes
Smo	smoothened homolog	8.17	10.45	6.3E-05	4.9	Mes
Gli1	GLI-Kruppel family member GLI1	7.78	10.44	1.6E-04	6.3	Mes
Gli2	GLI-Kruppel family member GLI2	9.39	11.17	1.3E-04	3.4	Mes
Gli3	GLI-Kruppel family member GLI3	6.73	9.94	4.1E-05	9.3	Mes
Hhip	Hedgehog-interacting protein	5.14	9.81	9.7E-04	25.4	Mes
Wnt Pathway						
Wnt4	wingless-related MMTV integration site 4	8.24	9.59	2.2E-03	2.5	Mes
Wnt5a	wingless-related MMTV integration site 5A	6.37	10.30	5.8E-05	15.3	Mes
Wnt8b	wingless related MMTV integration site 8b	9.00	8.00	5.5E-02	2.0	Epi
Wnt9a	wingless-type MMTV integration site 9A	6.82	7.93	2.7E-02	2.2	Mes
Fzd1	frizzled homolog 1	8.86	11.85	1.3E-04	8.0	Mes
Fzd2	frizzled homolog 2	6.84	10.80	6.8E-05	15.6	Mes
Fzd6	frizzled homolog 6	8.05	9.77	2.9E-03	3.3	Mes
Fzd7	frizzled homolog 7	6.81	10.21	2.4E-05	10.6	Mes
Dkk2	dickkopf homolog 2	5.88	9.21	3.8E-05	10.0	Mes
Dkk3	dickkopf homolog 3	6.95	9.96	1.5E-03	8.1	Mes
Sfrp1	secreted frizzled-related sequence protein 1	6.00	12.29	3.9E-06	78.5	Mes
Sfrp2	secreted frizzled-related sequence protein 2	6.49	10.04	7.7E-05	11.7	Mes
IGF Pathway						
Igf1	Insulin-like growth factor 1,	8.56	12.13	1.3E-05	11.9	Mes

	mRNA					
Igf1r	Insulin-like growth factor I receptor	7.64	10.71	5.2E-04	8.4	Mes
Igfbp2	insulin-like growth factor binding protein 2	7.49	11.98	2.2E-05	22.4	Mes
Igfbp3	insulin-like growth factor binding protein 3	5.72	12.40	4.1E-07	102.3	Mes
Igfbp4	insulin-like growth factor binding protein 4	8.00	13.41	9.0E-05	42.4	Mes
Igfbp5	insulin-like growth factor binding protein 5	6.22	11.55	7.4E-06	40.4	Mes
Igfbp6	insulin-like growth factor binding protein 6	8.14	10.60	7.5E-04	5.5	Mes
Igfbp7	insulin-like growth factor binding protein 7	7.79	13.01	5.4E-06	37.3	Mes
Igf2as	insulin-like growth factor 2, antisense	10.80	9.09	3.9E-03	3.3	Epi
Igf2bp3	insulin-like growth factor 2, binding protein 3	9.32	7.83	1.4E-02	2.8	Epi
FGF Pathway						
Fgf13	fibroblast growth factor 13	5.84	11.41	1.3E-05	47.6	Mes
Fgf15	fibroblast growth factor 15	9.53	8.00	6.8E-02	2.9	Epi
Fgf5	fibroblast growth factor 5	8.70	7.48	8.6E-03	2.3	Epi
Fgf7	fibroblast growth factor 7	7.20	8.34	2.9E-03	2.2	Mes
Fgfr1	fibroblast growth factor receptor 1	5.60	10.75	1.3E-05	35.5	Mes
Fgfr2	fibroblast growth factor receptor 2	9.31	11.20	2.0E-04	3.7	Mes
Spry2	sprouty homolog 2	7.85	10.69	1.1E-03	7.1	Epi
Spry1	sprouty homolog 1	9.05	11.38	2.8E-04	5.0	Mes
Spred2	sprouty-related, EVH1 domain containing 2	9.30	10.89	8.6E-03	3.0	Mes
Spred1	sprouty protein with EVH-1 domain 1	6.02	9.78	1.3E-03	13.6	Mes
Fgfbp1	fibroblast growth factor binding protein 1	10.62	7.99	2.6E-04	6.2	Mes

[a]Epi = Average expression value in epithelium (Log2) as calculated by RMA.
[b]Mes = Average expression value in mesenchyme (Log2) as calculated by RMA.
[c]ES = Enrichment score (absolute value of numerical fold difference, calculated as in Materials and Methods).
[d]E/M = Compartment in which the array data predict that the gene is expressed, epithelial (Epi) or mesenchymal (Mes).

Table 2.9 Compartmentalization Bmp pathway transcripts.

Symbol	Gene Title	Epi[a]	Mes[b]	P-value	ES[c]	E/M[d]
Bmp2	bone morphogenetic protein 2	8.74	9.87	2.8E-02	2.2	Mes
Bmp4	bone morphogenetic protein 4	7.79	12.36	5.5E-06	23.8	Mes
Bmp5	bone morphogenetic protein 5	6.57	11.22	8.6E-05	25.0	Mes
Bmp6	bone morphogenetic protein 6	7.95	9.66	4.4E-04	3.3	Mes
Bmp7	bone morphogenetic protein 7	10.57	8.33	4.1E-04	4.7	Epi
Bmpr1a	bone morphogenetic protein receptor, type 1A	8.42	9.89	1.7E-02	2.8	Mes
Bmpr2	bone morphogenic protein receptor, type II	10.44	12.44	7.7E-04	4.0	Mes
Acvr1	activin A receptor, type 1	8.14	10.32	3.5E-03	4.5	Mes
Smad1	MAD homolog 1	9.67	10.84	1.0E-02	2.3	Mes
Smad5	MAD homolog 5	9.37	11.02	1.4E-03	3.1	Mes
Smad6	MAD homolog 6	9.54	10.66	3.0E-02	2.2	Mes
Smad7	MAD homolog 7	9.67	10.70	1.8E-02	2.0	Mes
Bmp1	bone morphogenetic protein 1	6.61	8.71	3.7E-04	4.3	Mes
Twsg1	twisted gastrulation homolog 1	7.73	10.83	1.6E-03	8.6	Mes
Bmper	BMP-binding endothelial regulator	6.74	9.63	5.3E-04	7.4	Mes
Crim1	cysteine rich transmembrane BMP regulator 1	7.49	9.28	2.0E-02	3.5	Mes
Chrdl1	Chordin-like 1	5.38	8.05	7.9E-04	6.4	Mes
Tll1	tolloid-like	5.60	7.25	9.3E-04	3.1	Mes
Fstl1	follistatin-like 1	6.90	13.21	9.8E-08	79.3	Mes
Fst	follistatin	8.03	9.36	2.7E-03	2.5	Mes
Fstl5	follistatin-like 5	5.87	8.61	1.7E-04	6.7	Mes
Fstl3	follistatin-like 3	8.54	9.64	1.5E-03	2.2	Mes
Grem1	gremlin 1	8.01	11.22	1.4E-04	9.3	Mes
Sostdc1	sclerostin domain containing 1	6.11	8.71	1.9E-04	6.1	Mes
Gdf10	growth differentiation factor 10	7.22	10.34	6.2E-04	8.7	Mes

[a]Epi = Average expression value in epithelium (Log2) as calculated by RMA.
[b]Mes = Average expression value in mesenchyme (Log2) as calculated by RMA.
[c]ES = Enrichment score (absolute value of numerical fold difference, calculated as in Materials and Methods).
[d]E/M = Compartment in which the array data predict that the gene is expressed, epithelial (Epi) or mesenchymal (Mes).

BIBLIOGRAPHY

Beer HD, Bittner M, Niklaus G, Munding C, Max N, et al. 2005. The fibroblast growth factor binding protein is a novel interaction partner of FGF-7, FGF-10 and FGF-22 and regulates FGF activity: implications for epithelial repair. *Oncogene* 24: 5269-77

Buchwalter G, Gross C, Wasylyk B. 2005. The ternary complex factor Net regulates cell migration through inhibition of PAI-1 expression. *Mol Cell Biol* 25: 10853-62

Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, et al. 2005. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics* 21: 2933-42

Chambers D, Mason I. 2000. Expression of sprouty2 during early development of the chick embryo is coincident with known sites of FGF signalling. *Mech Dev* 91: 361-4

Chin KT, Zhou HJ, Wong CM, Lee JM, Chan CP, et al. 2005. The liver-enriched transcription factor CREB-H is a growth suppressor protein underexpressed in hepatocellular carcinoma. *Nucleic Acids Res* 33: 1859-73

Choi MY, Romer AI, Hu M, Lepourcelet M, Mechoor A, et al. 2006. A dynamic expression survey identifies transcription factors relevant in mouse digestive tract development. *Development* 133: 4119-29

Chuang PT, McMahon AP. 1999. Vertebrate Hedgehog signalling modulated by induction of a Hedgehog-binding protein. *Nature* 397: 617-21

Cohen P. 2006. Overview of the IGF-I system. *Horm Res* 65 Suppl 1: 3-8

Dessimoz J, Opoka R, Kordich JJ, Grapin-Botton A, Wells JM. 2006. FGF signaling is necessary for establishing gut tube domains along the anterior-posterior axis in vivo. *Mech Dev* 123: 42-55

Duluc I, Freund JN, Leberquier C, Kedinger M. 1994. Fetal endoderm primarily holds the temporal and positional information required for mammalian intestinal development. *J Cell Biol* 126: 211-21

Duluc I, Lorentz O, Fritsch C, Leberquier C, Kedinger M, Freund JN. 1997. Changing intestinal connective tissue interactions alters homeobox gene expression in epithelial cells. *J Cell Sci* 110 (Pt 11): 1317-24

Fraser JD, Martinez V, Straney R, Briggs MR. 1998. DNA binding and transcription activation specificity of hepatocyte nuclear factor 4. *Nucleic Acids Res* 26: 2702-7

Fre S, Huyghe M, Mourikis P, Robine S, Louvard D, Artavanis-Tsakonas S. 2005. Notch signals control the fate of immature progenitor cells in the intestine. *Nature* 435: 964-8

Garrison WD, Battle MA, Yang C, Kaestner KH, Sladek FM, Duncan SA. 2006. Hepatocyte nuclear factor 4alpha is essential for embryonic development of the mouse colon. *Gastroenterology* 130: 1207-20

Gray PA, Fu H, Luo P, Zhao Q, Yu J, et al. 2004. Mouse brain organization revealed through direct genome-scale TF expression analysis. *Science* 306: 2255-7

Green MC. 1968. Mechanism of the pleiotropic effects of the short-ear mutant gene in the mouse. *J Exp Zool* 167: 129-50

Gregorieff A, Pinto D, Begthel H, Destree O, Kielman M, Clevers H. 2005. Expression pattern of Wnt signaling components in the adult intestine. *Gastroenterology* 129: 626-38

Guler HP, Zapf J, Froesch ER. 1987. Short-term metabolic effects of recombinant human insulin-like growth factor I in healthy adults. *N Engl J Med* 317: 137-40

Haramis AP, Begthel H, van den Born M, van Es J, Jonkheer S, et al. 2004. De novo crypt formation and juvenile polyposis on BMP inhibition in mouse intestine. *Science* 303: 1684-6

He XC, Zhang J, Tong WG, Tawfik O, Ross J, et al. 2004. BMP signaling inhibits intestinal stem cell self-renewal through suppression of Wnt-beta-catenin signaling. *Nat Genet* 36: 1117-21

Hong J, Zhang G, Dong F, Rechler MM. 2002. Insulin-like growth factor (IGF)-binding protein-3 mutants that do not bind IGF-I or IGF-II stimulate apoptosis in human prostate cancer cells. *J Biol Chem* 277: 10489-97

Howe JR, Bair JL, Sayed MG, Anderson ME, Mitros FA, et al. 2001. Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis. *Nat Genet* 28: 184-7

Howe JR, Sayed MG, Ahmed AF, Ringold J, Larsen-Haidle J, et al. 2004. The prevalence of MADH4 and BMPR1A mutations in juvenile polyposis and absence of BMPR2, BMPR1B, and ACVR1 mutations. *J Med Genet* 41: 484-91

Hsu T, Trojanowska M, Watson DK. 2004. Ets proteins in biological control and cancer. *J Cell Biochem* 91: 896-903

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4: 249-64

Jacoby RF, Schlack S, Cole CE, Skarbek M, Harris C, Meisner LF. 1997. A juvenile polyposis tumor suppressor locus at 10q22 is deleted from nonepithelial cells in the lamina propria. *Gastroenterology* 112: 1398-403

Kas K, Finger E, Grall F, Gu X, Akbarali Y, et al. 2000. ESE-3, a novel member of an epithelium-specific ets transcription factor subfamily, demonstrates different target gene specificity from ESE-1. *J Biol Chem* 275: 2986-98

Kedinger M, Duluc I, Fritsch C, Lorentz O, Plateroti M, Freund JN. 1998. Intestinal epithelial-mesenchymal cell interactions. *Ann N Y Acad Sci* 859: 1-17

Lickert H, Kispert A, Kutsch S, Kemler R. 2001. Expression patterns of Wnt genes in mouse gut development. *Mech Dev* 105: 181-4

Madison BB, Braunstein K, Kuizon E, Portman K, Qiao XT, Gumucio DL. 2005. Epithelial hedgehog signals pattern the intestinal crypt-villus axis. *Development* 132: 279-89

McBride HJ, Fatke B, Fraser SE. 2003. Wnt signaling components in the chicken intestinal tract. *Dev Biol* 256: 18-33

Milano J, McKay J, Dagenais C, Foster-Brown L, Pognan F, et al. 2004. Modulation of notch processing by gamma-secretase inhibitors causes intestinal goblet cell metaplasia and induction of genes known to specify gut secretory lineage differentiation. *Toxicol Sci* 82: 341-58

Nakerakanti SS, Kapanadze B, Yamasaki M, Markiewicz M, Trojanowska M. 2006. Fli1 and Ets1 have distinct roles in connective tissue growth factor/CCN2 gene regulation and induction of the profibrotic gene program. *J Biol Chem* 281: 25259-69

Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, et al. 2004. Control of pancreas and liver gene expression by HNF transcription factors. *Science* 303: 1378-81

Ormestad M, Astorga J, Landgren H, Wang T, Johansson BR, et al. 2006. Foxf1 and Foxf2 control murine gut development by limiting mesenchymal Wnt signaling and promoting extracellular matrix production. *Development* 133: 833-43

Pabst O, Zweigerdt R, Arnold HH. 1999. Targeted disruption of the homeobox transcription factor Nkx2-3 in mice results in postnatal lethality and abnormal development of small intestine and spleen. *Development* 126: 2215-25

Parviz F, Matullo C, Garrison WD, Savatski L, Adamson JW, et al. 2003. Hepatocyte nuclear factor 4alpha controls the development of a hepatic epithelium and liver morphogenesis. *Nat Genet* 34: 292-6

Quandt K, Frech K, Karas H, Wingender E, Werner T. 1995. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* 23: 4878-84

Ramalho-Santos M, Melton DA, McMahon AP. 2000. Hedgehog signals regulate multiple aspects of gastrointestinal development. *Development* 127: 2763-72

Ratineau C, Duluc I, Pourreyron C, Kedinger M, Freund JN, Roche C. 2003. Endoderm- and mesenchyme-dependent commitment of the differentiated epithelial cell types in the developing intestine of rat. *Differentiation* 71: 163-9

Rehli M, Lichanska A, Cassady AI, Ostrowski MC, Hume DA. 1999. TFEC is a macrophage-restricted member of the microphthalmia-TFE subfamily of basic helix-loop-helix leucine zipper transcription factors. *J Immunol* 162: 1559-65

Simon-Assmann P, Kedinger M. 1993. Heterotypic cellular cooperation in gut morphogenesis and differentiation. *Semin Cell Biol* 4: 221-30

Stanger BZ, Datar R, Murtaugh LC, Melton DA. 2005. Direct regulation of intestinal fate by Notch. *Proc Natl Acad Sci U S A* 102: 12443-8

Stegmann A, Hansen M, Wang Y, Larsen JB, Lund LR, et al. 2006. Metabolome, transcriptome, and bioinformatic cis-element analyses point to HNF-4 as a central regulator of gene expression during enterocyte differentiation. *Physiol Genomics* 27: 141-55

Theodosiou NA, Tabin CJ. 2003. Wnt signaling during development of the gastrointestinal tract. *Dev Biol* 259: 258-71

van Es JH, van Gijn ME, Riccio O, van den Born M, Vooijs M, et al. 2005. Notch/gamma-secretase inhibition turns proliferative cells in intestinal crypts and adenomas into goblet cells. *Nature* 435: 959-63

Wahab NA, Mason RM. 2006. A Critical Look at Growth Factors and Epithelial-to-Mesenchymal Transition in the Adult Kidney. Interrelationships between Growth Factors That Regulate EMT in the Adult Kidney. *Nephron Exp Nephrol* 104: e129-e34

Wang CC, Biben C, Robb L, Nassir F, Barnett L, et al. 2000. Homeodomain factor Nkx2-3 controls regional expression of leukocyte homing coreceptor MAdCAM-1 in specialized endothelial cells of the viscera. *Dev Biol* 224: 152-67

Woodford-Richens K, Williamson J, Bevan S, Young J, Leggett B, et al. 2000. Allelic loss at SMAD4 in polyps from juvenile polyposis patients and use of fluorescence in situ hybridization to demonstrate clonal origin of the epithelium. *Cancer Res* 60: 2477-82

Yanagita M. 2005. BMP antagonists: their roles in development and involvement in pathophysiology. *Cytokine Growth Factor Rev* 16: 309-17

CHAPTER III

PATTERNING OF THE EPITHELIAL PYLORIC BORDER IS ACCOMPANIED BY A DRAMATIC INDUCTION OF GENE EXPRESSION IN THE INTESTINAL EPITHELIUM

Abstract

The embryonic stomach and intestine are generated from a continuous tube of endoderm wrapped in mesenchyme. Early in development, this bilayered tube receives an anterior/posterior pattern that marks out specific domains and controls organ shape, primarily by sculpting the mesenchymal component. In contrast, the epithelial component of the early tube is morphologically similar in stomach and intestine. However, in the adult, distinct morphological and transcriptional differences separate stomach epithelium from intestinal epithelium. Remarkably, the boundary between these two organs is literally one cell thick. We demonstrated earlier that, in the case of an intestine-specific gene (villin) this sharp junction is established suddenly and precisely at E16.5, by sharpening a previously diffuse anterior boundary of expression. In the present study, we define the dynamic transcriptome of stomach, pyloric border and intestinal tissues during formation of this boundary. We show that pyloric boundary formation is concomitant with an epithelial compartmentalization event that involves the simultaneous induction of about 2000 genes in the intestinal epithelium. Gene expression in the stomach, however,

changes little during this time. Intestinal genes that are induced are those that provide intestinal character; thus, we call this process intestinalization. We identify specific transcription factors and signaling proteins that are likely to play a role in this process. Finally, we identify genes that are expressed exclusively in the border region itself. This analysis reveals Gata3 and nephrocan as exciting new players in this important developmental event.

Introduction

The vertebrate gastrointestinal (GI) tract consists of a series of connected organs (esophagus, stomach, small intestine, large intestine), each with a highly specialized epithelial surface that enables it to perform its distinct function in digestion. In adults, the boundaries between adjacent organs are remarkably sharp. At the pylorus, for example, cells with all of the characteristics of stomach lie directly next to cells with intestinal character, and no intermediate cell type is detectable (Braunstein et al 2002).

These discrete organ boundaries have fetal origins. In the embryo, the gut tube is molded from endoderm, along with its associated splanchnic mesoderm (Wells & Melton 1999). Anterior/posterior patterning of the GI tract begins even before tube formation is complete. The developing gut tube has a clear Hox code that is detectable by E10 and marks out the major organ domains and future sphincter locations, with boundaries that appear sharp, at least by whole mount in situ hybrization (Kawazoe et al 2002). Expression patterns for other organ-specific transcription factors are also established early. For example, the HMG-box protein, Sox2, is expressed in the epithelium of the

stomach domain (Ishii et al 1998) while the caudal-related parahox factor, Cdx2, is expressed in the intestinal epithelium (Silberg et al 2000). In the mesenchyme, Nkx2.5 is expressed in a thin band at the site of the future pylorus that will divide the stomach and duodenal territories (Smith et al 2000, Theodosiou & Tabin 2005). These territorial domains may set the pre-pattern for the further maturation of the stomach/duodenum border. Remarkably, during this time of organ patterning, the endodermally-derived epithelial lining of the tube exhibits little obvious regional difference at the morphological level, even as late as E14.5.

In an earlier study, we pinpointed the exact developmental window during which a distinct epithelial border is formed between stomach and intestine (Braunstein et al 2002). We found that, at E14.5, the mouse villin gene, as well as a modified villin allele containing a β-galactosidase cDNA, is expressed at high levels in the intestine, with expression gradually diminishing across the presumptive pyloric boundary. However, at E16.5, a discrete boundary of expression is suddenly seen: villin (or β-gal in the villin$^{\beta\text{-gal}/+}$ model) is detectable at high levels in intestinal cells while neighboring stomach cells exhibit little or no expression (Braunstein et al 2002). This finding suggests that an important compartmentalization event occurs in the pyloric epithelium at E16.5. Interestingly, this late event occurs 5-6 days after the establishment of regional patterning in this region of the GI tube.

In the present study, we used microdissection and microarray analysis to examine gene expression in and around the pyloric border at E14.5 and E16.5, timepoints that

encompass the formation of the discrete stomach and duodenal compartments as defined by villin gene expression. The goal of these studies was to determine whether this apparent compartmentalization event is accompanied by similar expression changes in other genes. The results reveal that the villin gene is just one of about 2000 genes that are coordinately up-regulated in the intestinal epithelium at E16.5. We call this event, which clearly involves a burst of intestine-specific and epithelial cell-specific expression, *intestinalization*. The array reveals transcription factors (e.g., Hnf4γ, Tcfec) and key signaling pathways (e.g., Hedgehog and Wnt) that are dramatically modulated during this intestinalization event. Interestingly, we find that intestinalization also involves up-regulation of Creb3l3, a major determinant of the ER stress response pathway (Zhang et al 2006). The activation of this pathway may be necessitated by the large transcriptional burst that accompanies intestinalization. Finally, we show that several genes are restricted in their expression to the pyloric border itself; these include the previously identified genes, Nkx2.5 and gremlin (Moniot et al 2004), as well as new players, including Gata3 and the Tgfβ-inhibitor, nephrocan.

Materials and Methods

Dissection of tissues.

Embryos from C57Bl/6J mice were collected from timed pregnant females, with the day of vaginal plug detection considered day 0.5. Intestine and stomach were removed and three contiguous 1mm segments were collected from antrum, pyloric border and duodenum. The location of the pyloric border could be detected at both E14 and E16 as a constriction just distal to the stomach. Cuts were made on either side of this constriction. A total of 174 embryos were used for the E14 dissections and 95 embryos for the E16

dissections. Tissue was pooled in three separate pools for use in RNA extraction and cDNA preparation. Separate, independently collected pools were used in verification studies (RT-PCR).

Microarray experiment and normalization.

Tissue from each of six groups was homogenized in 1 mL of TRIzol (Invitrogen) and by drawing it through an 18-gauge needle. Total RNA was prepared according to the manufacturer's instructions and purified using the RNAeasy Kit (Qiagen). Samples were sent to the University of Michigan Cancer Center Microarray Core Facility for hybridization of labeled cRNA probe to Affymetrix MOE 430.2 arrays (eighteen total chips: 6 groups, 3 chips for each group). After chip hybridization, arrays were scanned to obtain the image files (.DAT files), which were then processed using MAS 5.0 software to produce .CEL files. The .CEL files were analyzed by the RMA method (Robust Multiarray Average, *affy* package in BioConductor), which subtracted background, normalized expression data and calculated log (base 2) gene expression values (Irizarry et al 2003b).

Microarray data analysis.

Expression values (Log2) were imported into MeV(Multi-experiment Viewer, one of TM4 tools designed by The Institute of Genome Research; http://www.tm4.org/) (Saeed et al 2006, Saeed et al 2003) for identification of statistically significantly changed genes using t-test and ANOVA. The list of statistically significant genes was selected based on p-values less than 0.05 and fold difference greater than 2.0 for further analysis. The fold change was calculated by subtracting the average Log2 expression value for one group

from the average Log2 expression values of another group (LogDif = Log2(GroupA) - Log2(GroupB). Then, the difference was converted to a numerical fold change (not log fold change) [$FC = 2^{(LogDif)}$].

Principal Component Analysis (PCA) was performed using functions and packages in R. For the chromosome analysis, the total number of genes in the genome and the number of genes on each chromosome were obtained from the NCBI database. The expected number of genes on particular chromosome was calculated by [(number of known genes on that chromosome)/(Total number of genes in the mouse genome)] x (the number of statistically significant genes). The information about chromosome locations for all genes on the candidate list (D16 upregulated epithelial genes) was downloaded from the Affymetrix database.

Transcription factor binding sites (TFBS) analysis was done using Genomatix tools and databases. The promoter sequences used were the default length of Genomatix (500bp upstream and 100bp downstream of the transcriptional start site or TSS). Pathway analysis was accomplished using GenMAPP, GeneGo, and BiblioSphere. GO term enrichment analysis and functional annotation in specific gene sets was also assessed using Genomatix (BiblioSphere) as well as the Database for Annotation Visualization and Integrated Discovery (DAVID) (http://david.abcc.ncifcrf.gov/) (Dennis et al 2003). Within DAVID, we utilized the newly implemented Functional Annotation tool. Given a list of genes, this tool uses clustering strategies to collate information from all three Gene Ontology categories (Cell Component, Biological Process and Metabolic Function) as well as from other sources (SwissProt, PIR, BioCarta, KEGG, etc.) in order to extract the most meaningful functional and pathway information (Huang da et al 2007).

In situ hybridization.

Staged E14.5 and E16.5 C57Bl/6J embryos were dissected in DEPC-treated 1X PBS. Isolated gastrointestinal tracts were fixed overnight at 4°C in 4% PFA, embedded in paraffin, and cut into 5 μm sections. In situ hybridization was performed as described previously (Li et al 2007). Specific probes used for in situ hybridization included: Gata3 (NM_008091.3; 1028-1591 bp), nephrocan (NM_025684.2; 758-1450 bp), Sfrp5 (NM_018780.2; 203-985 bp), Creb3l3 (NM_145365.3; 683-1361 bp), Tcfec (NM_031198.2; 258-789), Grem1 (NM_011824.4; 487-1281), and Axin2 (NM_015732.4; 2227-3358).

Antibodies and immunostaining.

Staged E14.5 and E16.5 C57Bl/6J embryos were dissected in 1X PBS. Excised gastrointestinal tracts were fixed for overnight at 4°C in 4% PFA, embedded in paraffin, and cut into 5 μm sections. Vectastain ABC (Vector) was used for Hnf4γ immunohistochemistry, per the manufacturer's instructions. Sox2 immunofluorescence was performed as described previously (Que et al., 2007). Briefly, sections were deparaffinized in xylene, rehydrated through increasing alcohol concentrations, and boiled for 10 minutes either in Antigen Unmasking Solution (Vector Laboratories, CA) (Sox2) or 10 mM sodium citrate, pH 6.0 (Hnf4γ). Slides were allowed to slowly cool down and then washed with 1X PBS. The sections were blocked with 10% donkey serum/0.01% Triton X-100 in 1X PBS for 1 hour at room temperature. Antibodies used were: rabbit polyclonal anti-Sox2 antibody (1:500, Chemicon, MA);

goat polyclonal anti-Hnf4γ (1:500, Santa Cruz Biotechnology, CA); biotinylated horse anti-goat IgG (1:200, Vector); and, Alexa Fluor 488 donkey anti-rabbit IgG (1:1000, Invitrogen, OR). Primary antibodies were diluted in blocking solution and incubated on slides overnight at 4^{o}C. Slides were washed in 1X PBS prior to incubation with appropriate secondary antibodies and DAPI (1:100, Fisher Scientific, PA) (Sox2) for 30 to 60 minutes at room temperature. After 1X PBS wash, coverslips were mounted with ProLong Gold Antifade Reagent (Invitrogen) (Sox2), or sections were lightly counterstained with hematoxylin, dehydrated with serial alcohol/xylene washes and coverslips were mounted with permount (Fisher). Imaging was done on a Nikon E800 microscope digital imaging system (Nikon, Japan).

X-gal staining.

Staged E14.5 embryos from genetic crosses of Gata3$^{LacZ/+}$ (kind gift of James Douglas Engel, University of Michigan) or Gli1$^{LacZ/+}$ (Park et al 2000) or Ptc1$^{LacZ/+}$ mice (Goodrich et al 1997) with C57Bl/6J mice were dissected on ice in 1X PBS. The gastrointestinal tract was excised, fixed with 4% paraformaldehyde (PFA) for 10 minutes at 4°C, and washed three times in 1X PBS. X-gal staining solution was prepared fresh, as follows: 1X PBS (pH 7.5), 2 mM magnesium chloride, 5 mM potassium ferrocyanide, 5 mM potassium ferricyanide, and 1 mg X-gal (from 20 mg/ml stock in DMF). Fixed tissue was incubated in staining solution at 37°C and monitored periodically to control staining intensity. Stained tissue was washed three times in 1X PBS and then post-fixed overnight at 4°C in 4% PFA. No background staining was observed for wild type embryos, even if the tissues were incubated overnight. Whole mount stained tissue was photographed in

1X PBS on a dissecting microscope (Leica, Germany) with a Spot CCD camera (Diagnostic Instruments, MI). For histological analysis, whole mount stained tissue was embedded in paraffin, cut into 5 μm sections, and stained lightly with eosin. Slides were photographed on a E800 microscope (Nikon) with a Spot CCD camera.

RT-PCR.

Samples from four independent collections were separated by tissue type and time point (e.g. - E14.5 border, E16.5 duodenum, etc.) and then pooled randomly into two groups for replicate analysis. Additionally, fragments of contiguous tissue spanning the antrum, border and proximal duodenum were collected from E14.5 and E16.5 embryos as an input control. Tissue from each group was homogenized in 1 mL of TRIzol (Invitrogen) by drawing it through a 30-gauge needle. Total RNA was prepared according to the manufacturer's instructions, purified by consecutive phenol-chloroform (2X) and chloroform (2X) extractions, and quantitated by UV spectrophotometry using a NanoDrop (Thermo). For each RNA sample, two independent cDNA preparations were performed using the iScript cDNA Synthesis Kit (Bio-Rad) and pooled for subsequent analysis. Negative "No RT" controls for genomic contamination were prepared from whole E14.5 and E16.5 RNA in a similar manner. PCR was performed on cDNA samples using qRT-PCR-quality primers created by Beacon Designer (PREMIER Biosoft). Individual primer conditions were optimized (i.e.- temperature, number of cycles, magnesium chloride, DMSO, etc; see Table 3.1) prior to PCR. Products from PCR reactions were resolved under UV light with ethidium bromide-loaded, 2% TBE-agarose

gels. The band intensity of experimental genes was compared to the housekeeping gene *Hprt*.

Results

Late development of the epithelial pyloric boundary.

Expression of Sox2 in the developing stomach has been previously demonstrated by whole mount in situ hybridization; the boundary of the staining domain at the pylorus is quite distinct, even by E12. However, to understand the process of epithelial pyloric border formation, it was important to determine whether, at the cellular level, the early expression domain of this patterning protein is characterized by a sharp, one cell thick boundary at the pylorus or by a fading gradient of expression as seen for the villin gene. The appearance of villin expression, as reflected by a modified villin$^{\beta\text{-gal}/+}$ allele (Madison et al 2002), at E14.5 is shown in Figure 3.1A. Similarly, the expression domain of Sox2 is, like villin, diffuse at E14.5 (Figure 3.1B). Thus, it appears that while a regional GI epithelial pattern is established before E14.5, the pyloric boundary of this pattern is not mature at the cellular level until E16.5. Figure 3.1C demonstrates this for villin$^{LacZ/+}$ staining.

Epithelial border formation is accompanied by changes in the global transcriptome of stomach and duodenum.

To learn more about the process of epithelial pyloric border formation, we analyzed gene expression on both sides of the developing border and at the border itself at two time

points: E14.5 (before border formation) and E16.5 (after border establishment). The processes for harvesting and dissection of tissues and for microarray analysis are described in Methods. Individual pools of RNA were used to hybridize three Affymetrix chips for each time and tissue studied (two times x three tissues x 3 chips = 18 chips). The range of signal intensity in the original and normalized data is shown in Figure 3.2.

Principal component analysis was done to assess the similarities and differences among the various samples (Figure 3.3A). This analysis shows that the three chips that represent each of the six temporal and spatial groups are clustered, indicating that the sampling is reproducible. In addition, the three E14.5 samples are grouped while the three E16.5 samples are clearly different from one another and different from the E14.5 groups. This indicates that at E14.5, stomach, border and duodenal tissues are similar to one another, but at E16.5, the three tissues are dramatically different. Among the three tissues sampled, the duodenum shows the most change, as measured by the degree of separation between E14.5 duodenum and E16.5 duodenum along the x axis; the x axis represents Principal Component 1, the component that contains the majority of the variation in the data. In contrast, the stomach samples exhibit much less change along the x coordinate.

Gene expression changes between E14.5 and E16.5 are primarily duodenal.

A list of all pair-wise expression changes with fold change ≥ 2.0 and $p \leq 0.05$ along time and spatial dimensions was generated and is available on the Gut group Coursetools website (http://ctools.umich.edu/portal/). This Master Table includes pairwise comparisons of: a) stomach, intestine and border to one another at E14.5 and at E16.5 and

b) each tissue at E14.5 to the same tissue at E16.5. A total of 12,445 changes (considering time and tissue) were detected that meet the criteria above. It is not possible because of space constraints to list all of these data in this thesis; therefore, I will outline the major findings and show specific datasets where they are relevant to the conclusions.

A summary of all results for pairwise comparisons along the time axis is shown in the ven diagram in Figure 3.3B. Of the 9181 probesets that change over time, only 1069 exhibit change in all three tissues. The duodenum exhibits the most robust change over time, with differences in a total of 7787 probesets, 4976 or 64% of which change only in the duodenum and not in border or stomach tissues. In the stomach, 3057 probesets show temporal change but just 28% of these are stomach-specific. The border tissue exhibits temporal change in 2548 transcripts, only 8% of which are unique to the border region. Thus, while transcriptional change is detectable in all three tissues between E14.5 and E16.5, the transcriptome of the duodenum clearly demonstrates the greatest degree of modification.

We also tallied transcriptional changes that are specific to both space and time. For example, an E14.5 stomach-specific group was collated from those probesets that show significant change (upregulated or downregulated) in the E14.6 to E16.5 comparison (S14 to S16) AND significant change between duodenum and stomach at E14.5 (S14 to D14). In this manner, E14.5 stomach-specific (S14), E16.5 stomach-specific (S16), E14.5 duodenum specific (D14) and E16.5 duodenum-specific (D16) groups were defined. The results of those tallies (Figure 3.3C) reveal that at E14.5, the number of probesets that

exhibit stomach-specific changes in expression (304) exceeds those showing duodenum-specific changes (143). By E16.5, the stomach-specific set increases 1.8 fold, but the duodenum-specific set shows a dramatic, 24-fold increase. These results clearly mirror the results of the principal components analysis, showing that stomach and duodenum exhibit few differences from one another at E14.5, but are robustly different at E16.5. The vast majority of the change that occurs over this time period occurs in the duodenum. Tables listing all changes specific to each group are available on the Gut group web site (http://ctools.umich.edu/portal/).

Upregulated genes in the duodenum are primarily epithelial; down-regulated genes are mesenchymal.

In an earlier study, we separated intestinal tissue using non-enzymatic methods and profiled gene expression in isolated epithelium and mesenchyme to create a catalog of epithelially-expressed and mesenchymally-expressed genes (Li et al 2007). Though the earlier study was done using E18.5 intestine, we reasoned that the major epithelial/mesenchymal compartmentalization of genes would be similar at E16.5 and E18.5. Thus, using this compartmentalization catalog, we tagged all D16-specific genes as epithelial (1347 genes) or mesenchymal (1234 genes). Another 830 genes were expressed in both compartments and thus could not be classified in this way. We then analyzed the 3411 D16-specific genes, separating them into depleted (D14 > D16, 1420 genes) or enriched (D14 < D16, 1991 genes). Applying the epithelial/mesenchymal designations to these separated sets yielded dramatic results. For D16-enriched genes that could be classified as epithelial or mesenchymal, a total of 1344 of 1366 (98%) were

epithelial (Figure 3.2D). In contrast, for D16-depleted genes that could be cataloged in this way, 1212 of 1215 (99.7%) were mesenchymal.

Table 3.2 records fold changes measured for the D16 group among enriched and depleted genes. A total of 323 of the enriched genes (16%) were up-regulated by 10 fold or greater (29 of these went up over 50 fold), while only 13 of the depleted genes (0.9%) were down-regulated by greater than 10 fold. Thus, the change in duodenal gene expression that accompanies establishment of the epithelial pyloric border is characterized by robust up-regulation of gene expression in the epithelium and less dramatic but still significant down-regulation of gene expression in the mesenchyme.

Validation of microarray results.

The array results suggested that pyloric border formation involves impressive changes in gene expression that result in the evolution of very different gene expression patterns in duodenum and stomach. To provide independent validation for the array results, a few genes were selected from each temporal/special-specific group and primers were generated to examine mRNA expression using 28-30 cycles of PCR. These results, shown in Figure 3.4, reveal 100% concordance with the array findings and emphasize the dramatic differences in expression that characterize the stomach and duodenum.

Chromosomal location of epithelial genes.

Since our analysis suggested a coordinate up-regulation of over 1000 genes specifically in the intestinal epithelium, we examined the chromosomal localization of these genes to

determine whether this might reflect localized and coordinated transcriptional activity of clusters of genes on one or a few chromosomes. Though this analysis did reveal enrichment of up-regulated genes on Chr. 9 and 11 (Figure 3.5), these genes did not map to clusters, but were relatively distributed on the chromosomes (data not shown). Interestingly, there was a relative paucity of D16 up-regulated epithelial genes on the X chromosome (Figure 3.5).

Gene Ontology analysis of enriched and depleted D16-specific genes.

We used the functional annotation clustering tool in the DAVID resource at http://david.abcc.ncifcrf.gov/ to functionally classify the genes that are modulated trancriptionally at D16. Table 3.3 shows that genes involved in metabolic processes and biosynthesis are the most robustly affected set in the D16-enriched genes in the epithelium. Interestingly, though few genes are up-regulated in D16 mesenchyme (22), these genes are largely involved in the immune response and inflammatory processes (Table 3.3). Among genes depleted in the mesenchyme at D16, those contributing to extracellular matrix and cell adhesion are most frequently represented. Depleted genes in D16 epithelium numbered only three (glucagon, Foxa1 and melanoregulin); thus functional annotation clustering could not be applied to this group. However, since Foxa1 is a transcription factor, this result could have functional significance. The results of this analysis indicate that the large burst of gene expression in the intestine at D16 is preparing the intestine for its major role in both metabolism and immunity.

Expression of transcription factors during epithelial pyloric border formation.

Transcription factors were identified by Gene Ontology tags as described earlier (Li et al 2007). Table 3.4A-D lists the most enriched transcription factors for each of the four temporal and spatial groups described above (e.g., D16-specific transcription factors are those that are more than two fold different in the D14 to D16 comparison AND the D16 to S16 comparison, etc.). Note that fold change values are most robust for the D16 group, where the dynamic range for the D16 to D14 comparison is 40 fold and the range for the D16 to S16 comparison is up to 26 fold. In contrast, most other groups exhibit fold change values less than 10. Thus, by this measure again, the D16.5 duodenum is undergoing particularly robust changes in gene expression.

Modulation of signaling pathways during epithelial pyloric border formation.

Soluble signaling factors play major roles in gut patterning, and previous data suggest that Bmp (Moniot et al 2004, Smith et al 2000), Hh (Zhang et al 2001a) and Wnt (Kim et al 2007a) signaling may be important in the context of pyloric border formation and/or intestinal differentiation. Thus, we examined the array results for expression of key elements of these pathways. The Fgf pathway was also examined, as this pathway plays an important role in formation of the midbrain-hindbrain boundary (Canning et al 2007, Scholpp et al 2003, Trokovic et al 2005). These results are summarized in Table 3.5.

Several Fgf family members (Fgf 1, 13 and 14), two receptors (Fgfr1 and Fgfr2) and two intracellular inhibitors (Spry1 and Spred1) were dynamically expressed in these tissues. Expression levels of all proteins were similar in stomach and duodenum at E14. However,

Fgfbp1 was nearly 8 fold up-regulated upon stomach differentiation and many of the other family members (except notably Fgf1) were somewhat down-regulated with differentiation of the duodenum. Thus, by E16, Fgf1 was more prominent in the duodenum, while Fgf13 and Fgf14 were 4-6 fold more prominent in the stomach.

Only one of the four Notch family members was found to be dynamically regulated in this study. Notch2 showed down-regulation during intestinal development so that at E16.5, expression was 3 fold greater in the stomach. Notch expression exhibited a mesenchymal pattern in intestine.

The Hedgehog family exhibited the most consistent response of all signaling families examined. At E14.5, all family members were expressed similarly in stomach and intestine (Table 3.5). However, at E16.5, Shh, which is expressed in the epithelium and signals in a paracrine manner to the mesenchyme (Madison et al 2005), was down-regulated 9 fold in the duodenum only. All three Gli factors were reduced 3-6 fold in the duodenum, as were Gas1 and Boc, two recently identified co-receptors (Allen et al 2007). The down-regulation of all of these proteins specifically in the duodenum and not in the stomach leaves a robust gradient of Hh pathway activity across the pyloric border at E16.5.

Among genes of the Wnt pathway, regulation is also dynamic (Table 3.5). Sfrp5 is prominent in the duodenum at both E14.5 and E16.5; duodenal expression is 13-30 fold greater than stomach at E14.5 and up to 10 fold greater than stomach at E16.5.

The downregulation of Sfrp1 and Sfrp2 in stomach at E16.5, noted in an earlier study (Kim et al 2005a), is also revealed here; in addition, we see an even more robust down-regulation of these genes in the duodenum (Table 3.5). Also noteworthy are the Wnt receptors Fzd1 and Fzd2, which are expressed at similar levels in E14 stomach and duodenum and are not modulated during stomach development over this time period. However, both messages are down-regulated in differentiating duodenum, so that by E16, at least for Fzd2, stomach expression is four fold greater than that of duodenum.

Earlier investigations demonstrated that Bmp signaling is directly involved in the formation of the pyloric border (Moniot et al 2004, Smith et al 2000, Theodosiou & Tabin 2005). Thus, this family was of particular interest. Only one Bmp protein, Bmp7, was found to be differentially regulated during pyloric border formation (Table 3.5). At E14.5, Bmp7 was expressed at similar levels in stomach and duodenum, but it was up-regulated during differentiation in the duodenum (not in stomach) so that by E16, duodenal expression was up to 3 fold greater than stomach expression. The Bmpr1b receptor showed the opposite pattern; it was down-regulated with duodenal differentiation, so that at E16.5, it was more prominent in the stomach. This agrees well with an earlier study (Kim et al 2007b). Both Id4 (a Bmp target gene) and Twsg1, a modulator of Bmp activity, were regulated similarly to Bmpr1b, suggesting that overall Bmp signaling is greater in the stomach than in the duodenum at E16.5. A similar pattern was seen for multiple components of the Igf signaling pathway. In general, all of these signaling pathways are notable (with very few exceptions: Igfbp7 and Sfrp1) for their lack of dynamic regulation during maturation of the stomach from E14.5 to E16.5 (Table

3.5). In contrast, differentiation of the duodenum is characterized by numerous robust changes in the mRNAs encoding these signaling proteins, the vast majority of which are expressed in the mesenchyme.

Dynamic gene expression in the duodenum is coordinate with formation of a sharp epithelial boundary at the pylorus.

The data above establish that the D14 to D16 transition is accompanied by a large burst of gene expression in the duodenum. We looked carefully at the boundaries of expression of several genes at the E14 and E16 pyloric border to determine whether these genes would be regulated similarly to villin and Sox2 above. If so, we would expect a sharp, single cell thick boundary at E16. Both immunostaining and *in situ* hybridization was used for these studies.

Hnf4γ, Tcfec and Creb3l3 are three of the most dynamically regulated transcription factors in the D16 group. Figure 3.6 shows that, indeed, all three of these factors exhibit a a sharp epithelial boundary at the pylorus at E16.5. Details about the compartmentalization of these factors are additionally revealed. For example, immunhistochemical staining for Hnf4γ reveals both nuclear and cytoplasmic staining in the epithelium. At E16.5, cytoplasmic staining extends into the stomach, but nuclear staining is sharply demarcated at the pyloric border (Figure 3.6A-B). For both Tcfec (Figure 3.6C-E) and Creb3l3 (3.6F-G), expression is restricted to only the differentiated cells of the villus tips. It is likely that Hnf4γ, Tcfec and Creb3l3 are at least in part, responsible for the dramatic up-regulation of many metabolic genes in the E16 epithelium

(see Discussion). In fact, we earlier demonstrated that a large number of epithelial-specific genes of the duodenum have binding sites for Hnf4 in their promoters (Li et al 2007).

Our analysis of signaling factors above, collated in Table 3.5, establishes that the entire Hedgehog pathway is robustly down-regulated in the E16 duodenum and that the soluble Wnt inhibitor, Sfrp5, exhibits the most dramatic regulation of any signaling factor examined. It was important to confirm both of these findings experimentally. To examine the difference in Hedgehog signaling at E16, we utilized $Gli1^{LacZ/+}$ reporter mice, in which a β-galactosidase cDNA is incorporated at the endogenous Gli1 locus and faithfully reports Gli1 activity (Park et al 2000). Since Gli1 is a direct Hedgehog target, this also measures pathway activity. Figure 3.7A shows that at E14.5, $Gli1^{LacZ/+}$ staining, and thus Hh pathway activity, is similar in stomach and intestine. However, at E16.5, Gli1 activity is reduced on the duodenal side of the pyloric border, but maintained in the stomach, in agreement with the array results. This difference is also visible in whole mount stained stomach and intestine from $Ptch1^{Lac/+}$ mice (Figure 3.7C), another reporter of Hh activity (Goodrich et al 1997). This difference in pathway activity is likely a direct consequence of the significant down-regulation of Shh in the duodenum from E14 to E16 (Table 3.5).

Sfrp5 expression was examined using *in situ* hybridization. Interestingly, Sfrp5 is highly expressed in the duodenal epithelium at E14, with a soft anterior boundary of expression that extends a short distance into the stomach (Figure 3.8A). A survey of additional

regions of the gut tube reveals that Sfrp5 is duodenum-specific at this time (figure 3.8B, arrow). The D14-S14 expression difference is 13-31 fold in the array (Table 3.5). By D16, when villus formation has begun, this expression domain resolves dramatically; Sfrp5 expression is excluded from the villus tips and present only in a few cells of the proliferative, intervillus region (Figure 3.8C,D).

Sfrps are soluble regulators of Wnt expression and the pattern of Sfrp expression suggested that Wnt signaling may be modulated across the pylorus. Interestingly, Sfrp5 seems to be the only Wnt modulator that is expressed at a higher level in the duodenum than in the stomach. Sfrp1, Sfrp2 and Sfrp4 as well as Dkk2 and Dkk3 seem to be expressed at similar levels in stomach and duodenum at E14, at time when Sfrp5 exhibits up to 30 fold concentration in the duodenum (Table 3.5). At E16, though Sfrp5 is down-regulated in the duodenum, it retains its preference for the duodenal domain. In contrast, the other factors are all downregulated 2-4 fold in both stomach and duodenum. To establish whether a gradient of Wnt signaling indeed exists across the pylorus, we examined the expression of Axin2, a Wnt target gene and a commonly used read-out of Wnt signaling activity (Wang et al 2008, Yan et al 2001). Figure 3.8C-D shows the Axin2 pattern, as determined by in situ hybridization. Strikingly, the pattern suggests that at E14.5, Wnt signaling is similar in stomach and duodenum (Figure 3.9A). At this time, the Axin2 signal is greater in distal intestine than it is in the duodenum (Figure 3.9A, arrow, Figure 3.9B). A closer look at this distal staining reveals that cells closer to the epithelial basement membrane are experiencing higher levels of Wnt signaling than the luminal cells (Figure 3.9B). At E16, the activity of this pathway is clearly detectable in

the duodenum (Figure 3.9D,E). Importantly, Axin2 staining is exclusive to the intervillus region of E16 intestine, contradicting a recent study in which the region of active Wnt signaling in intestine at E16.5 was proposed include only the villus tip epithelium (Kim et al 2007a).

A specific domain of gene expression at the pyloric border.

Earlier studies had revealed that, in the chick, Nkx2.5 expression is restricted to the mesenchyme surrounding the pyloric border where, under control of Bmp signaling, it interacts with the gizzard protein, Sox 9, to direct the formation this border (Moniot et al 2004, Smith et al 2000, Theodosiou & Tabin 2005). Our analysis revealed that a number of genes are expressed in the border tissue at levels >2 fold greater than in surrounding stomach or duodenum. Fifteen probesets were enriched or depleted at the border at E14.5; 79 were detected at E16.5. These genes are listed in Table 3.6A and 3.6B.

Only one gene, melanoregulin, was found to be depleted in the border tissue relative to stomach (3 fold) and duodenum (3 fold) at E14.5. Melanoregulin purportedly plays a role in organelle biosynthesis by regulating peripherin-regulated membrane fusion (Boesze-Battaglia et al 2007); the reason for its exclusion from the border region at E14.5 is not clear.

Table 3.6A confirms earlier data showing enrichment of both Nkx2.5 (5-6 fold relative to surrounding tissues) and gremlin, a modulator of Bmp signaling (7 fold relative to

stomach and 2 fold relative to duodenum) at E14.5. Novel genes expressed in this domain include: Lgals6 (5-7 fold enriched), a galectin family member; Lect1 (4 fold enriched), also known as chondromodulin-I, an angiogenesis inhibitor (Hiraki & Shukunami 2005); Skiv2l2 (3.8 fold enriched), a DEAD-box RNA helicase that may play a role in proliferation (Yang et al 2007); Arrdc3 (2.5 fold enriched), a tumor suppressor protein that is induced by Vitamin D3 and is a regulator of cell cycle progression (Mandalaywala et al 2008); Usp3 (2 fold enriched), a chromatin modifier and regulator of ubiquitination (Nicassio et al 2007); and Skap2 (enriched 2 fold), a src family kinase that inhibits PTK2B/RAFTK activity (Harada et al 2008). The most impressive enrichment detected in the border tissue at E14.5 was for Gata3 (7-10 fold), a zinc finger transcription factor that plays roles in multiple developmental processes (Ho & Pai 2007).

Table 3.6B lists 78 genes that are regulated specifically at the border at E16.5. The most dramatic changes are seen in the following genes. Nephrocan is 16-17 fold enriched in border vs. surrounding tissues. Nephrocan, which is also highly expressed in the kidney, is a member of the leucine-rich repeat family and a potential regulator of Tgfβ activity (Mochida et al 2006). Another urothelial protein, uroplakin 1A (4-5 fold enriched in border tissue) is an apical membrane protein in urothelium (Veranic et al 2004), while Slc4a1 (enriched 4.5 to 6.5 fold) is a chloride/bicarbonate exchanger that is important for acid/base balance in the kidney (Stehberger et al 2007). Procollagen, type IX, alpha 1, or Col9a1 is a collagen gene that, interestingly, is required for heart valve development (Lincoln et al 2006); it is enriched 4-10 fold at the pyloric border.

Because the role of gremlin had been earlier inferred in pyloric border maturation (Moniot et al 2004), we examined changes in its expression using in situ hybridization. Figure 3.10A-C shows that this gene is mesenchymal and is expressed in a pattern similar to that seen earlier for Nkx2.5 (Smith et al 2000, Theodosiou & Tabin 2005). From E14.5 to E16.5, the expression of gremlin decreases only slightly. In addtion to its concentration at the pylorus, gremlin is expressed in the inner circular muscle of both antrum and intestine (Figure 3.10A-C). Another secreted factor, nephrocan, is expressed in pyloric epithelium (Figure 3.10D-F). Its expression domain is slightly asymmetric with respect to the border, with higher expression on the stomach side of the developing pylorus. At E16.5, nephrocan becomes restricted to cells at the intervillus base of the intestine and cells at the base of developing antral glands. There is no clear boundary of expression across the pylorus for nephrocan (Figure 3.10E-F).

The novel findng of Gata3 concentration at the pyloric border at E14.5 and E16.5 (Table 3.6A and 3.6B) was also verified by additional studies. By serendipity, another laboratory in our department, that of Dr. J. Douglas Engel, had already generated a modified allele of Gata3, containing a β-galactosidase marker (unpublished data). Using this mouse model, we examined LacZ expression during embryonic development. We found a distinct and well-demarcated band of LacZ expressing cells surrounding the pyloric region at all stages examined, E12.5 to E16.5 (Figure 3.11A-E). The band of staining was open to the ventral side (Figure 3.11C). Furthermore, a spur of cells seems to extend back from the LacZ positive region, towards the stomach (Figure 3.11D, arrow and 3.11G). In sectioned material, LacZ staining was present in the mesenchyme in the

periphery of the sectioned pylorus (Figure 311E-G). Thus, the Gata3 expression domain in the pyloric mesenchyme seems to partially overlap with that of gremlin and Nkx2.5.

Discussion

In this study, we have investigated dynamic gene expression patterns in the developing pylorus. Interestingly, though clear regional patterning is present at E14, this involves a relatively small number of genes. In general, the transcriptomes of stomach, duodenum and border tissues are surprisingly similar at E14, as is best illustrated in the Principal Components Analysis in Figure 3.3A. In dramatic contrast, at D16.5, a burst of robust transcriptional induction involving about 2000 genes occurs the duodenum, but not in the stomach. We show that this genetic induction event takes place in the intestinal epithelium, not in the mesenchyme; in fact, transcription within the mesenchyme is largely reduced. The purpose of this transcriptional burst is clearly to activate intestine-specific epithelial genes involved in absorption and metabolism. As a result of this induction, and the morphological changes that accompany it, the differentiated, functional compartment of the intestinal epithelium is generated. Since the intestinal tissue has changed from a ground state characterized by few differences with stomach to a state characterized by the expression of hundreds of genes in intestine, but not stomach, we call this process “intestinalization”.

Though the vast majority of genes that are up-regulated in the intestine are epithelial, a few (22) mesenchymal genes are also up-regulated. It is interesting that these genes are predominantly involved in immune or inflammatory function. Recent parallel evidence

from our laboratory lead us to speculate that the dramatic down-regulation of Hh signaling that we also detect in the intestine might play a role in the initiation of this inflammatory signature in the mesenchyme (Lees et al 2008). We recently demonstrated that murine myeloid and dendritic cells in the lamina propria are cellular targets of Hh and that decreased Hh signal transduction leads to increased susceptibility to dextran sodium sulfate-induced colitis and increased inflammatory signaling (Lees et al., 2008). We also identified a non-synonymous SNP in the 3' end of the Gli1 molecule that results in a sub-functional transactivator. This change, E1100Q, is associated, in a dose-specific manner, with inflammatory bowel disease in three large European populations (Lees et al., 2008). Whether there is indeed a direct connection between the down-regulation of the Hh pathway and this activation of mesenchymal inflammatory genes is an important target for future study. Interestingly, among robustly activated epithelial genes is caspase-1, the major pro-inflammatory caspase that is responsible for the processing of Il-1β. Thus, both the epithelium and the mesenchyme seem to be preparing for a prominent role in mucosal immune function.

Several epithelial transcription factors are dramatically up-regulated and may participate directly in this large-scale induction of the absorptive and metabolic activity of the intestine. The Hnfγ paralog Hnf4α has been previously implicated in a similar developmentally late maturation event in the liver. In that system, Hnf4α up-regulates a large number of structural genes and is thought to be important for the epithelialization of hepatic cells following their migration out of the gut tube proper and into the septum transversum (Parviz et al 2003). It is tempting to speculate that intestinalization is a

similar event. Even though the intestinal epithelial cells never leave the confines of the epithelial sheet, as developing hepatoblasts do, the epithelium itself is drastically remodeled during the process of villus formation. Perhaps Hnf4α and Hnf4γ are important in the final stabilization of the remodeled state. Certainly, as documented in the previous chapter, binding sites for these factors are highly represented in the promoters of intestine-specific genes (Li et al 2007).

Another transcription factor that is very highly induced in the E16.5 intestinal epithelium is Tcfec. This bHLH factor is a member of the Mitf family (Tcfeb, Tcfec, Tcfe3). Several of the proteins in this family are expressed in a highly tissue- and cell-specific manner. Typically, these proteins are responsible for the expression of signature cell-specific proteins that are important in cellular development and function. For example, they regulate tartrate-resistant alkaline phosphatase in osteoclasts (Partington et al 2004), melanin in pigment cells (Tachibana 2000) and mast cell proteases in mast cells (Nechushtan & Razin 2002). These proteins form both homodimers and heterodimers, a fact that may explain the fact that the knockouts of several family members show no phenotypes despite the apparent transcriptional importance of these genes. Tcfeb and Tcfe3 are also expressed in the intestine, but neither of these proteins is regulated in the dramatic manner that Tcfec is during intestinalization.

Creb3l3 is a member of the bZip family of transcription factors and is involved in the ER stress response. It is of interest that Creb3l3 is regulated similarly in the liver (late induction) and is a known transcriptional target of Hnf4α (Luebke-Wheeler et al 2008).

Since intestinalization is an event that involves the transcriptional activation of over 1000 genes, several of which are highly expressed, it is possible that the rough endoplasmic reticulum must suddenly attain a much higher degree of organization and efficiency to deal with the translational onslaught to follow. Indeed, we show that Creb3l3 is expressed in villus epithelial cells, exactly the population in which differentiated gene expression is induced. The idea that an ER stress response might accompany development is novel, but has been well documented for cell maturation (e.g. in the case of the plasma cell) (Wu & Kaufman 2006).

Intestinalization occurs concomitantly with the formation of a sharp boundary of gene expression in the epithelium of the pyloric border. For genes like villin and Sox2, a broad domain of expression is detectable early and reflects the regional divisions of territory in the developing gut tube. But at E16.5, the boundary of expression sharpens exquisitely to allow the differentiated intestinal cells to lie directly next to future stomach cells. An interesting question for further analysis is whether border formation and intestinalization are under separate or simultaneous control. That is, does the process of intestinalization need to stop somewhere, and does this result in a discrete boundary? If so, what regulates the location of this boundary? Or, does the boundary become specified, propagating a signal that promotes intestinalization, similar to the function of a classical organizer? In this regard, it is interesting that pyloric border patterning is similar in some ways to the formation of the midbrain-hindbrain border of the brain and the atrioventricular boundary of the heart. Both of these events, like the establishment of the pyloric junction, involve formation of straight, sharp expression boundaries (Canning et al 2007, Chi et al 2008,

Joyner et al 2000). In both cases, the border region itself has signaling activity that influences neighboring tissues (Bai et al 2002, Chi et al 2008). Though organizer activity has not yet been demonstrated for the pyloric border it is noteworthy that at least two signaling proteins are expressed there: gremlin and nephrocan.

The intestinalization event that we have identified occurs without a similar maturation event in the stomach; the number of genes that show stomach specific expression at E16.5 (553) is less than double the number that exhibits stomach-specific expression at E14.5 (304). In contrast, by E16.5, the intestine has increased the number of compartment-specific genes by 24 fold (from 123 to 3411). The fact that the intestine, a more posterior tissue, matures prior to the stomach is somewhat surprising given the tendency of embryonic development to progress in an anterior to posterior direction. Indeed, it is possible that this finding has evolutionary roots. The stomach is believed to be an added character that first appeared in primitive fish. A hypothesis that fits both with the more recent evolution of the stomach and with the findings of our study would be that a repression program is in place in the stomach domain, the purpose of which is to prohibit the stomach from invoking an intestinal character during this time, so that it could later be given instructions to become stomach. Though this notion is entirely speculative, it has interesting implications for intestinal metaplasia, a pathological lesion in which patches of epithelium with intestinal character emerge in the stomach. The possibility that active repression of intestinal character exists during pyloric border formation and persists in adult life will become amenable to further investigation now

that the transcriptomes of stomach and intestine are available during this important developmental event.

Acknowledgements

The authors thank Dr. Alex Joyner for providing the $Gli1^{LacZ/+}$ mice. D.G. acknowledges support from R01 DK065850; A.U. is grateful for support from F30 DK082144. We thank experts in the Michigan Cancer Center Microarray Core Facility for help with the hybridization and scanning of chips; we also acknowledge support from the Microscopy and Image Analysis Laboratory of the Department of Cell and Developmental Biology as well as the Morphology Core of the Center for Organogenesis.

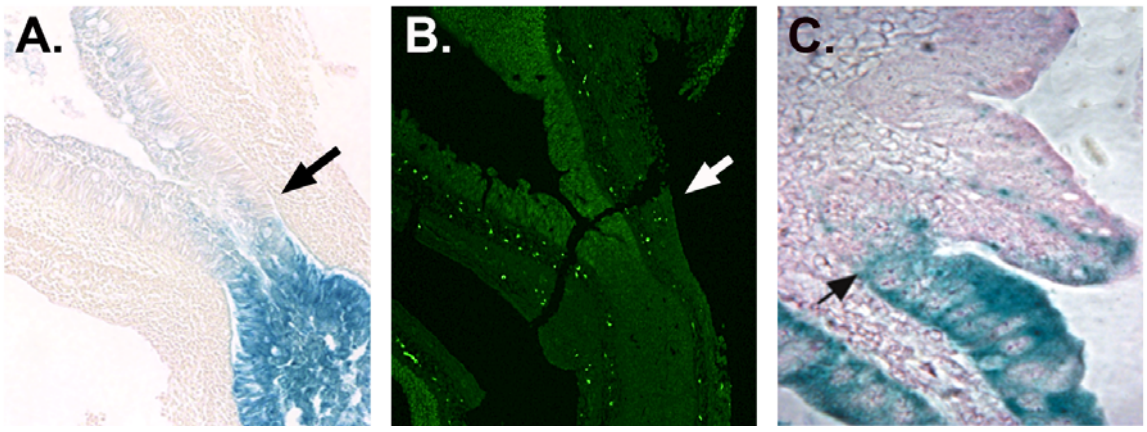

Figure 3.1 The epithelial pyloric boundary is diffuse at E14.5 for villin, an intestinal structural gene and for Sox2, a transcription factor expressed in the stomach. A) LacZ expression at E14.5 is shown for a modified allele of the villin locus that carries a β-galactosidase cDNA (Braunstein et al., 2001). Stomach is to the left of the arrow, intestine is to the right. B) Immunostaining for Sox2, an HMG family transcription factor that has been implicated in stomach patterning. Stomach is to the left of the arrow, intestine is to the right. The section is from an E14.5 embryo. C) Formation of a distinct boundary of villin expression in the pyloric epithelium at E16.5. Stomach is to the left of the arrow, intestine is to the right. By this time, Sox2 expression has regressed in the anterior direction; it no longer stains the pyloric region to an appreciable degree.

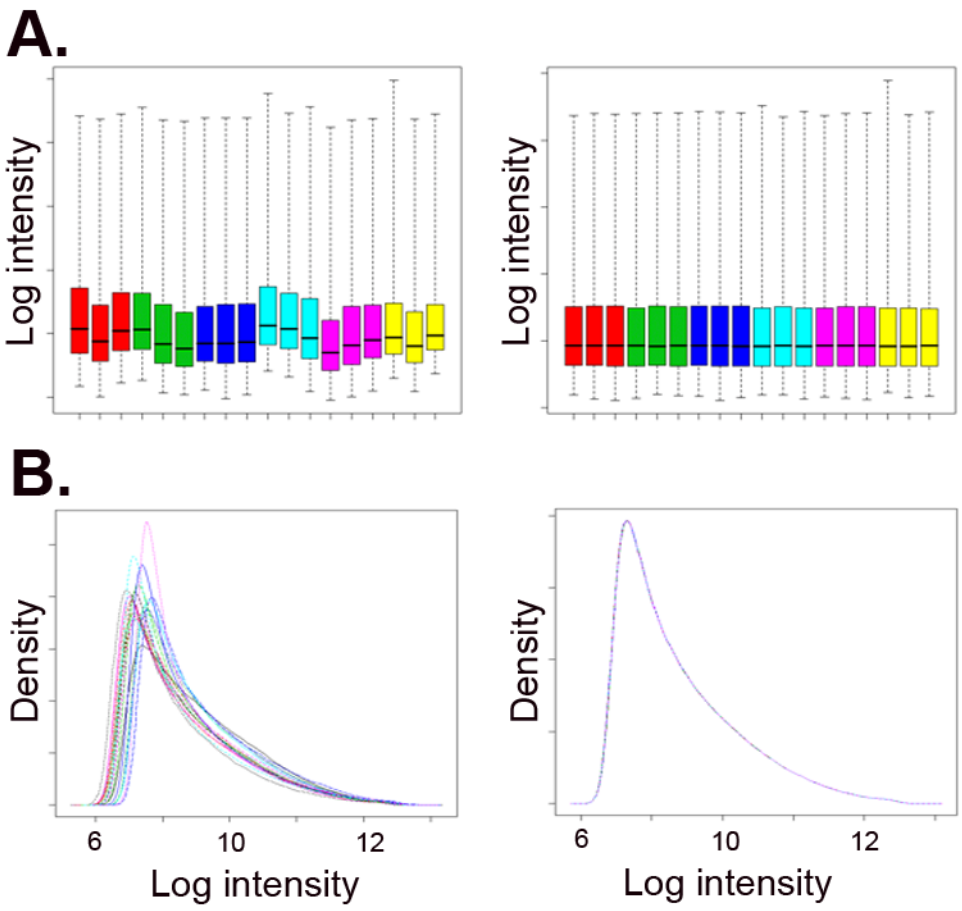

Figure 3.2 Microarray data normalization and distribution. The top two panels are box plots describing the range of signal intensity (y axis) for all chips in the array (x axis). The left top graph describes the original data in the .CEL file, while the right top graph represents the same data after normalization by RMA. The bottom figures display the range of values seen in the array. Again, the un-normalized data is to the left and the data after normalization is on the right.

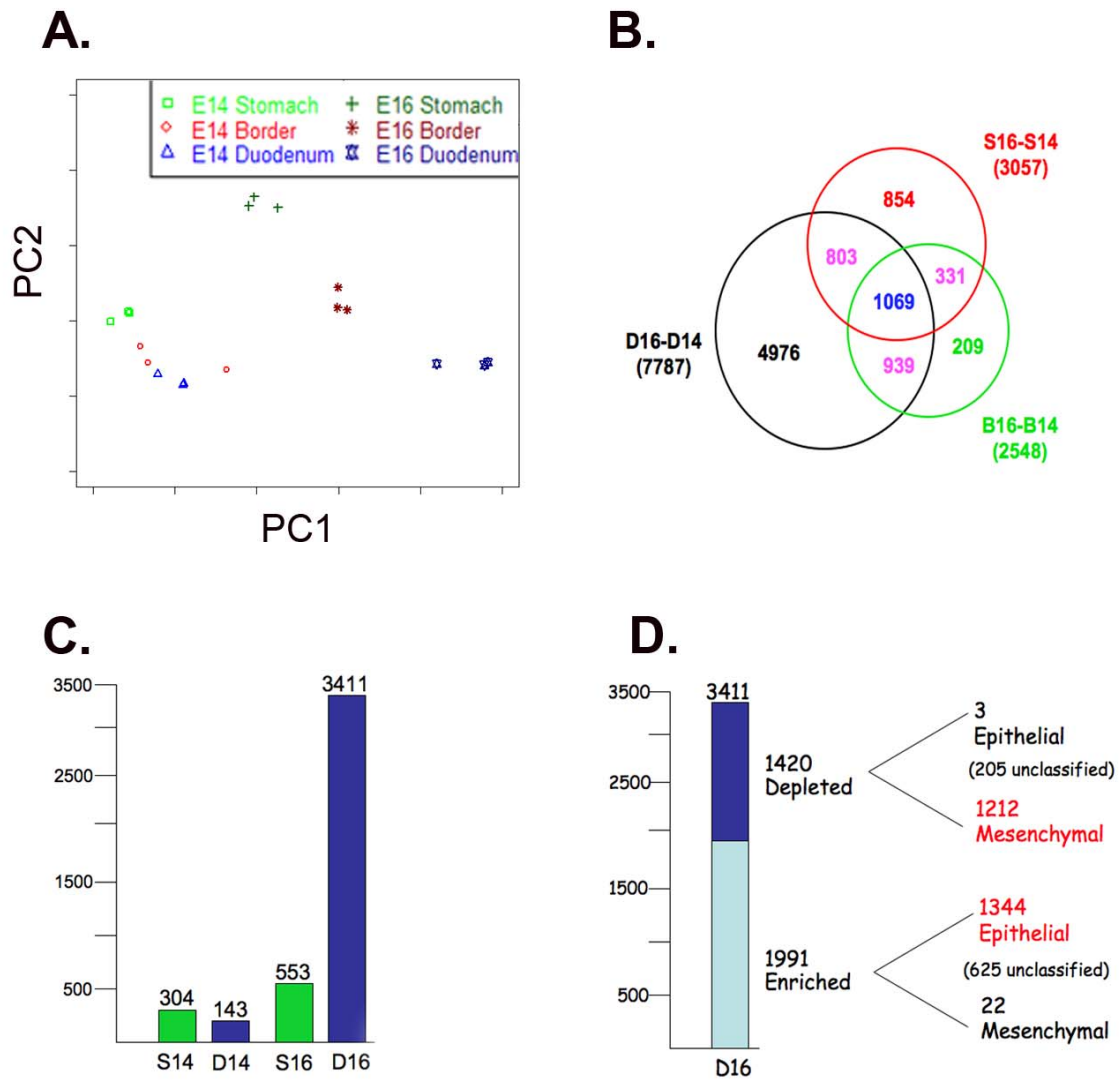

Figure 3.3 Overview of microarray results. A) Principal Components Analysis. The first two principal components, which together represent the majority of the variation in the data, are plotted here. The results are discussed in the text. B) Ven diagram showing overlaps in pairwise comparisons over the time axis. The three circles indicate all genes that differ >2.0 fold and $p \leq 0.05$ in duodenum (black), stomach (red) and border (green). Some genes fit these criteria in more than one tissue. These are shown by the overlaps. C) Number of genes that differ along the time AND tissue axis. For example, S14 genes differ when compared to S16 AND D14. See text for details. D) Compartmentalization (epithelial vs. mesenchymal) of enriched and depleted genes from the D16 group shown in (C).

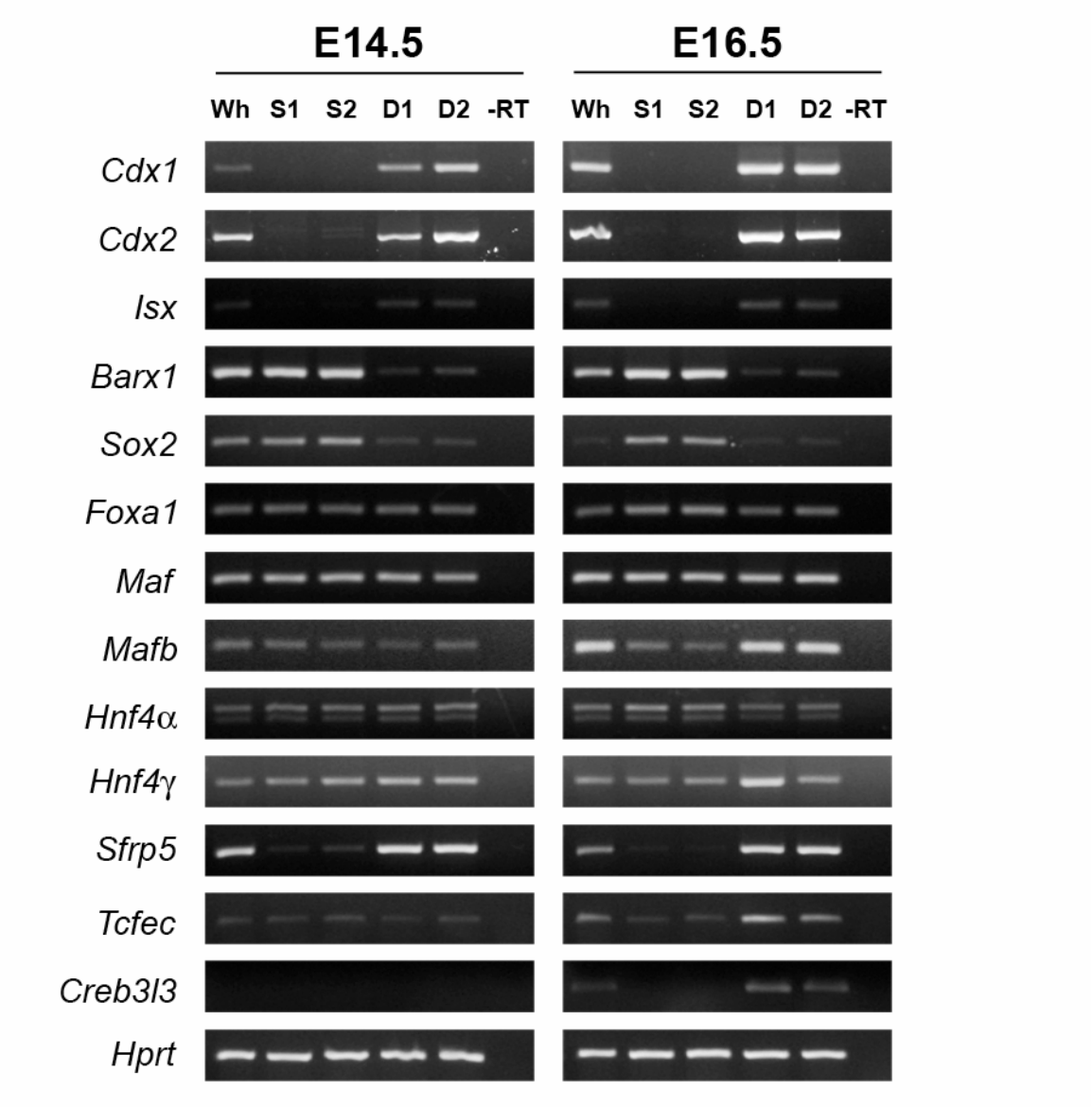

Figure 3.4 RT-PCR verification of some of the genes found to be time and tissue specific in the microarray. RT-PCR was carried out as described in Methods. Hprt is used as a control; it does not vary with time and tissue. Tissue specific regulation of Cdx1, Cdx2, Isx, Barx1 and Sox2 is set by E14.5 and does not vary with time; these represent genes responsible for pre-pattern of the intestine and stomach. Mafb and Hnf4γ are preferentially, but not specifically, expressed in the intestine. Finally, Tcfec and Creb3l3 are greatly induced in E16.5 duodenum.

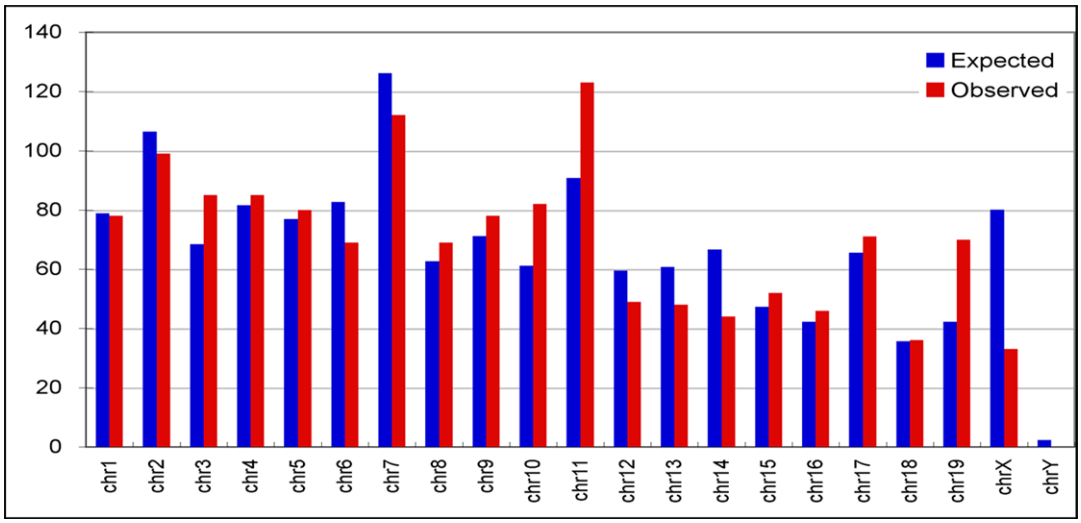

Figure 3.5 Chromosomal location of all genes that are upregulated in the D16 epithelium. The Expected number (blue bars) was calculated as the relative number of genes on each chromosome (# of genes on the chromosome/total genes x total genes in the upregulated D16 epithelial group). The red bars indicate the number of D16 up-regulated epithelial genes that are actually mapped to each chromosome.

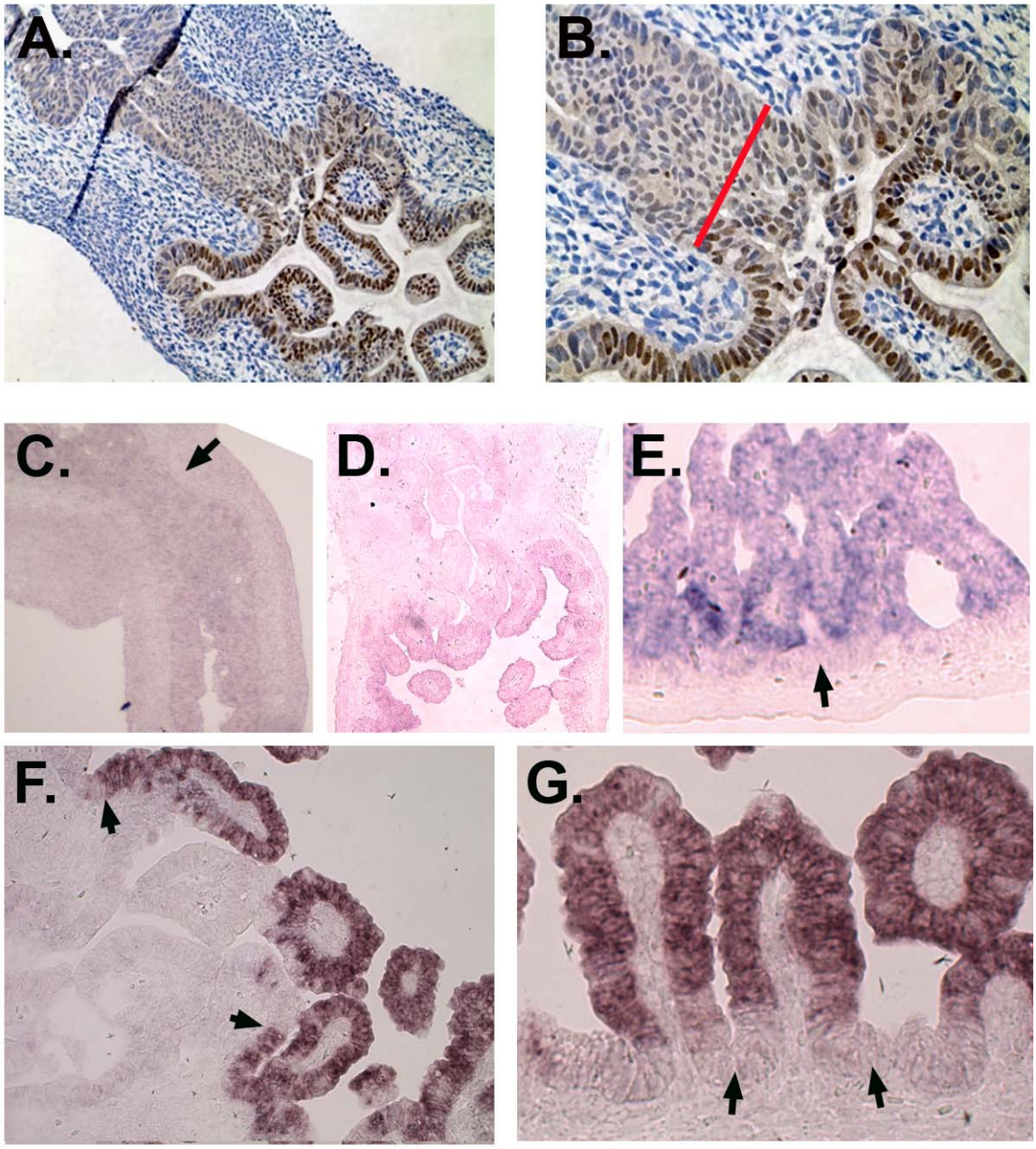

Figure 3.6 Verification of expression patterns for Hnf4γ, Tcfec and Creb3l3. A-B) Immunohistochemical staining for Hnf4γ shows nuclear staining in the intestine and cytoplasmic staining in the stomach (top), with a sharp boundary between the two patterns at the pylorus. A higher power view is shown in B. C-E). Tcfec is not expressed appreciably at E14.5 (C), but at E16.5, it is expressed in differentiated duodenal villus epithelium. A sharp boundary of staining is visible at the pylorus (D). No staining is seen in the intervillus area (arrow, E). F-G) Creb3l3 is not expressed at E14.5 (data not shown), but is expressed in villus epithelium, not intervillus epithelium (arrows in F,G). A sharp boundary of expression is seen at the pylorus (F).

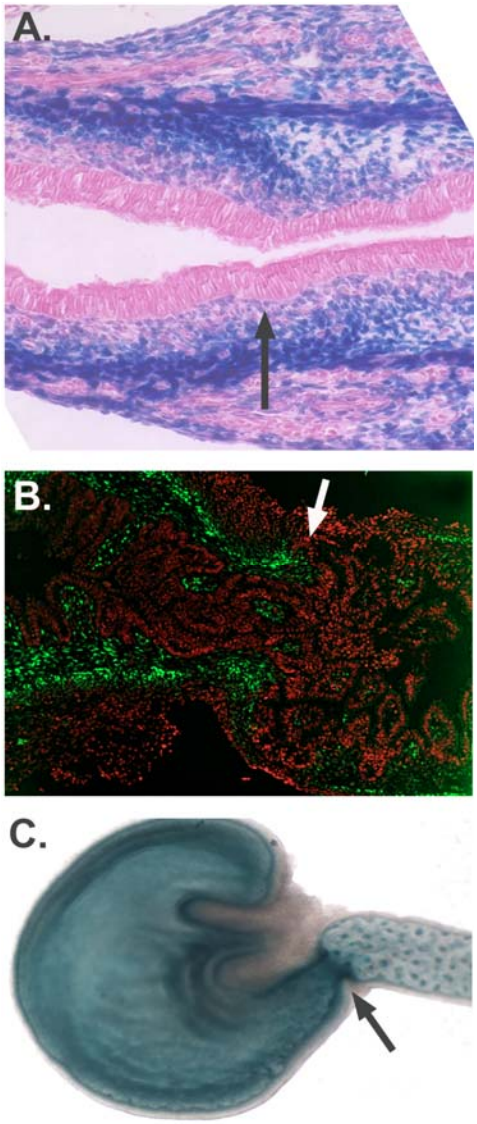

Figure 3.7 Hedgehog signaling decreases across the pyloric border at E16.5. A) section of the pyloric region from an E14.5 $Gli1^{LacZ/+}$ reporter mouse. Signaling is similar on both sides of the presumptive pyloric border (arrow). Intestine is to the right of the arrow; stomach is to the left. B) E16.5 $Gli1^{LacZ/+}$ reporter mouse. Red staining is DAPI; Green staining is anti-β-gal (antibody kindly provided by J.Douglas Engel, University of Michigan Department of Cell and Developmental Biology). The white arrow indicates the pyloric border. Intestinal tissue to the right of the arrow has much less β-gal signal (green) than does stomach tissue to the right of the arrow. C) E16.5 Ptch1 LacZ/+ mouse stomach and intestine stained with X-gal. A dramatic staining gradient is visible across the pylorus (arrow).

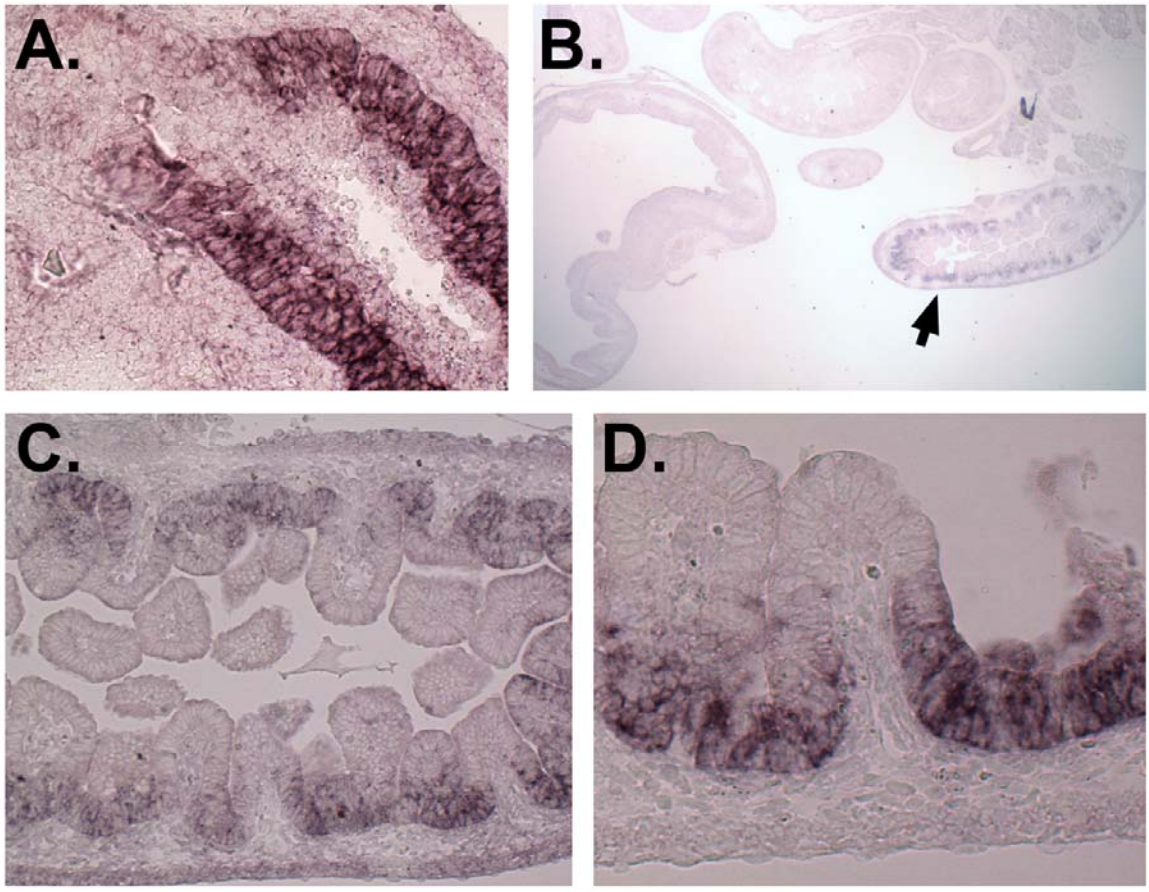

Figure 3.8 Sfrp5 expression at E14.5 and E16.5. A) Section across the pyloric border at E14.5; stomach is to the left, intestine is to the right. Note the gradual fading of Sfrp5 expression across the epithelial border. B) Survey of Sfrp5 expression at E16.5. The stomach is negative for Sfrp5, as is the distal intestine. However, the proximal intestine is positive (arrow). C, D) Higher magnification images of Sfrp5 expression in the E16.5 intestine. Expression is strong in the intervillus regions and little expression is seen on villus tips.

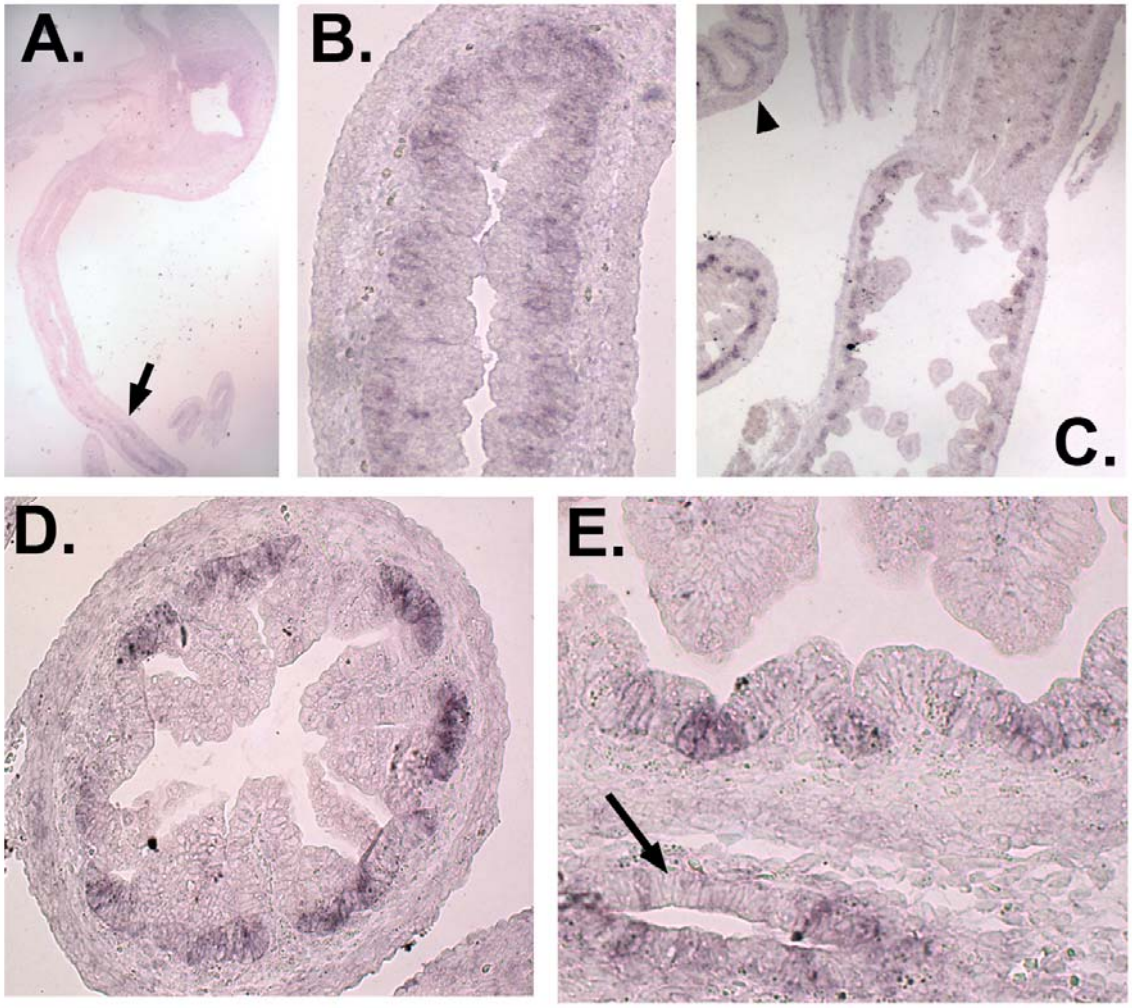

Figure 3.9 Expression of Axin2 in E14.5 and E16.5 stomach and intestine. A) E14.5 stomach and intestine. Very little expression of Axin2 is seen in stomach and proximal duodenum; faint staining is visible in the epithelium of the proximal jejunum (arrow). B) Higher magnification image of the area indicated by the arrow in (A). Note that the multilayered E14.5 epithelium is positive for Axin2 only in the region next to the basement membrane of the epithelium. Luminal cells are negative. C) At 16.5, faint staining is seen in both the stomach and duodenal epithelium, with little difference in staining between these two tissues. Staining is also visible in the forstomach (arrowhead). D) Higher magnification image of the intestine. Staining is restricted to the intervillus epithelium. E) Staining in the intervillus epithelium of the duodenum and in the pancreatic duct (arrow).

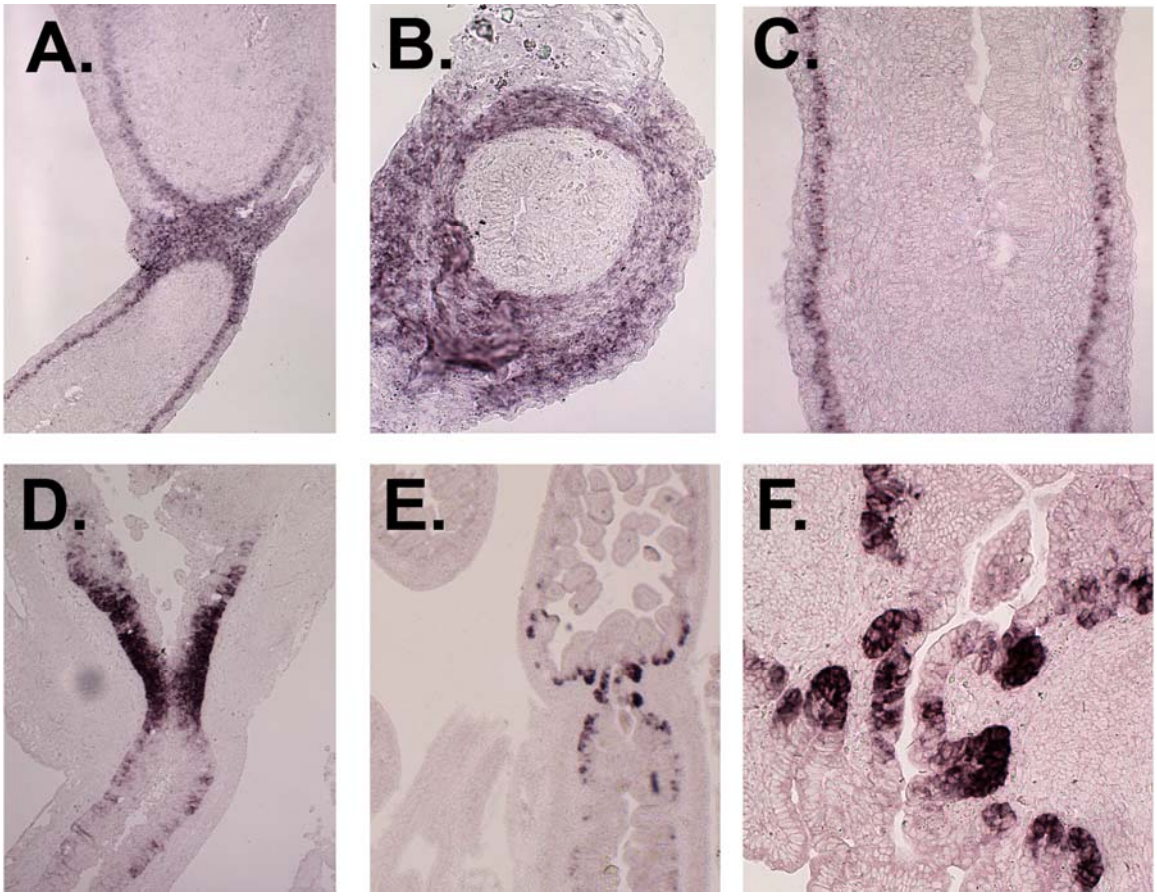

Figure 3.10 Border-specific expression of two secreted signaling proteins: gremlin in the mesenchyme and nephrocan in the epithelium. A) E14.5 pyloric border, showing gremlin expression in the mesenchyme. Expression continues in the inner circular muscle of the intestine and more weakly in the inner circular muscle of the stomach. B) Cross-section of the pyloric border at E16.5, demonstrating robust gremlin expression in elongated cells of the mesenchyme. C) At E14.5, gremlin is expressed throughout the intestinal inner circular muscle. D) Expression of nephrocan in pyloric epithelium at E14.5. Note that expression is more robust on the stomach side of the border. In both stomach (top) and intestine (bottom), expression is only seen in epithelial cells closest to the basement membrane. E) Nephrocan expression at the pyloric border at E16.5 is restricted to the base of the developing epithelial glands in stomach and intestine. This is the only portion of the gut that expresses nephrocan. F) Higher magnification of nephrocan expression at the pyloric border. Stomach is at the top of the image, intestine is at the bottom. Note that expression is most intense at the base of developing glands in the stomach and in the intervillus regions of the intestine. Discontinuous staining is visible across the epithelial pyloric border.

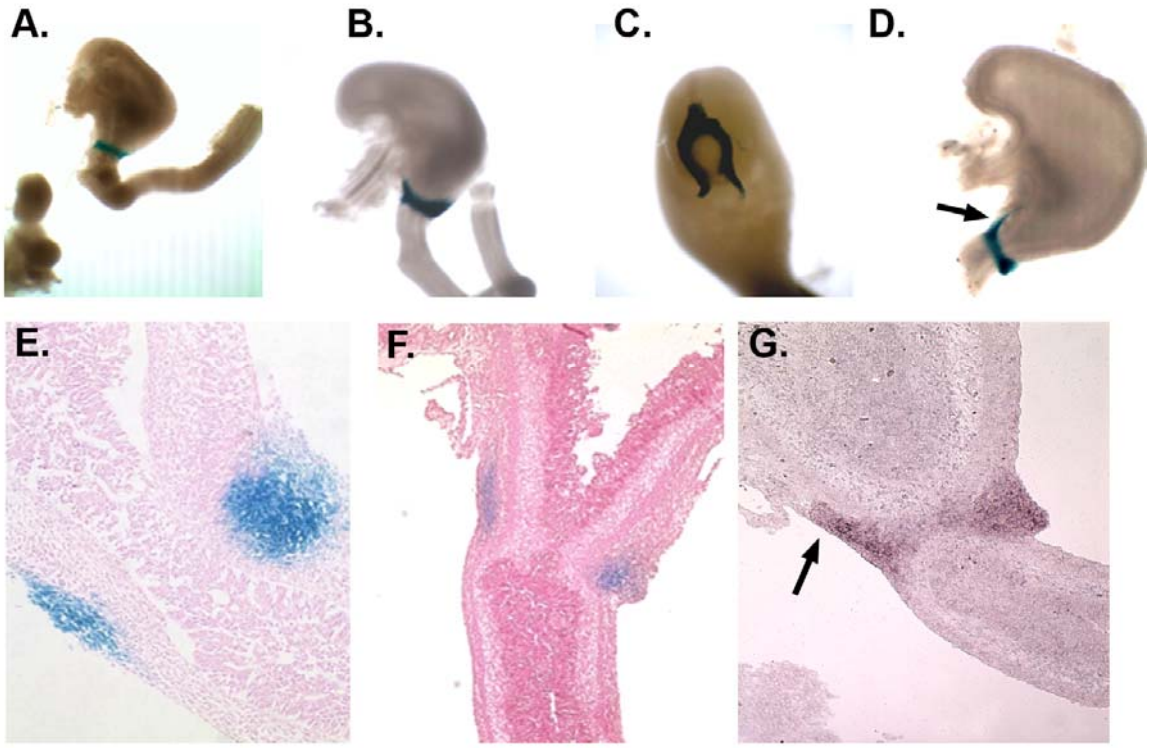

Figure 3.11 Expression of Gata3 at the pyloric border. A) E12.5 stomach from a Gata3 LacZ/+ knockin mouse model shows exquisite patterning of Gata3 expression at the pyloric border. B) E13.5 stomach; Gata3 expression continues at the pylorus. C) In a stomach with the duodenum removed, it is obvious that the Gata3 staining pattern is saddle-shaped and discontinuous on the ventral (lesser curvature) side. A "horn" of the saddle is present on the dorsal (greater curvature) side. D) E14.5 embryo. Arrow denotes a spur of Gata3 expression that extends up the ventral side toward the esophagus. E) Section from a LacZ stained E14.5 embryo reveals that β-gal expression is restricted to the mesenchyme. The inner circular muscle is not stained. It is not clear if the stained cells are muscle cells. G) E16.5 embryo shows a similar staining pattern. G) In situ hybridization using a Gata3 antisense probe shows mesenchymal staining at the pylorus and confirms the existence of the ventral spur of staining. These results are entirely consistent with the data seen using the Gata3-LacZ knockin model.

Table 3.1 RT-PCR primer sequence and optimized conditions.

Gene Symbol	Forward primer	Reverse primer	Amplicon length (bp)	Temp. (deg C)	Cycles	$MgCl_2$ (mM)	DMSO[1]
Barx1	GGAGTCGCACCGTATTCACTGAGC	CCGCCACCTTGCAGCACTATTTTC	187	58	30	2.5	NO
Cdx1	CGGACGCCCTACGAATGGATG	CTGCTGCTGCTGCTGTTTCTTC	276	61	30	1.5	YES
Cdx2	GAAACCTGTGCGAGTGGATGC	TTGTTGCTGCTGCTGCTGTTG	284	58	30	1.5	YES
Creb3l3	CAGTGGCATCTCTGAGGATCTACC	CAGTGAGGTTGAAGCGGGAGG	266	61	30	1.5	YES
Foxa1	GGCATGAGAGCAACGACTGG	TAGGTGTTCATGGAGTTCATAGAGC	115	57	30	2.5	NO
Hnf4a	GATGCTTCTCGGAGGGTCTGC	GATGCTTCTCGGAGGGTCTGC	200	60	30	1.5	NO
Hnf4g	CGCAGCATTCGGAAGAGTCATG	CCGCTTGTGCCAGAGTGTTTATG	220	60	30	1.5	NO
Hprt	AGTCCCAGCGTCGTGATTAGC	ATAGCCCCCCTTGAGCACACAG	204	61	30	2	YES
Isx	ACTTCACCCATTACCCTGACATCC	TCTTCTCCTGCTTCCTCCACTTG	123	55	30	1.5	YES
Maf	CAACGGCTTCCGAGAAAACG	CCCACGGAGCATTTAACAAGG	111	56	30	2.5	YES
Mafb	GCAACAGCTACCCACTAGCC	AGCTGGTCATCAGAGAAGCG	108	60	30	2.5	YES
Sfrp5	CCCTGGACAACGACCTCTGC	CACAAAGTCACTGGAGCACATCTG	143	59	30	2.5	YES
Sox2	CTCGCAGACCTACATGAACG	AGTGGGAGGAAGAGGTAACC	146	59	28	3	NO
Tcfec	ATGAACCCATGAGCCCAGACAG	AGCATCCGTGAGACCAGCATTAG	173	56	30	2	YES

[1]2% final concentration

Table 3.2 Fold changes (D16/D14) in the set of D16 enriched and depleted probesets

	Total Probesets	100 fold	50 fold	10 fold	5 fold	3 fold
D16 Enriched	1991	6	29	323	695	941
D16 Depleted	1420	0	0	13	256	972

This table examines probesets in the D16-specific group (see Table 3). It examines specifically the D16 to D14 change and tallies the number of probesets that experience various degrees of change (from 100 fold to 3 fold) over that time, for both up-regulated (enriched) or down-regulated (depleted) probesets. The enriched group has many more highly-regulated probesets than the depleted group.

Table 3.3 Functional Annotation Clusters Analysis by DAVID

D16-Enriched in epithelium

Annotation Cluster 1	Enrichment Score: 6.81			Count	P_Value	Benjamini
GOTERM_BP_ALL	organic acid metabolic process	RT		50	1.7E-14	2.9E-11
GOTERM_BP_ALL	carboxylic acid metabolic process	RT		49	6.4E-14	8.3E-11
GOTERM_BP_ALL	monocarboxylic acid metabolic process	RT		29	3.6E-11	3.8E-8
SP_PIR_KEYWORDS	lipid metabolism	RT		19	2.3E-9	9.9E-7
GOTERM_BP_ALL	fatty acid metabolic process	RT		20	2.2E-7	1.2E-4
GOTERM_BP_ALL	nitrogen compound metabolic process	RT		32	5.3E-7	2.5E-4
SP_PIR_KEYWORDS	Fatty acid metabolism	RT		11	1.1E-6	1.6E-4
GOTERM_BP_ALL	amino acid biosynthetic process	RT		10	2.6E-6	1.1E-3
GOTERM_BP_ALL	glutamine family amino acid metabolic process	RT		9	4.9E-6	1.7E-3
KEGG_PATHWAY	Fatty acid metabolism	RT		11	9.1E-6	4.5E-4
GOTERM_BP_ALL	amine metabolic process	RT		28	1.1E-5	3.0E-3
GOTERM_BP_ALL	nitrogen compound biosynthetic process	RT		12	3.3E-5	6.1E-3
GOTERM_BP_ALL	amino acid and derivative metabolic process	RT		24	4.6E-5	7.9E-3
GOTERM_BP_ALL	amine biosynthetic process	RT		10	1.1E-4	1.8E-2
GOTERM_BP_ALL	amino acid metabolic process	RT		19	3.1E-4	3.7E-2
Annotation Cluster 2	**Enrichment Score: 5.36**			**Count**	**P_Value**	**Benjamini**
SP_PIR_KEYWORDS	oxidoreductase	RT		44	2.4E-8	5.3E-6
GOTERM_MF_ALL	oxidoreductase activity	RT		56	6.7E-7	3.0E-4
GOTERM_BP_ALL	electron transport	RT		24	4.9E-3	3.3E-1

D16-Enriched in mesenchyme

Annotation Cluster 1	Enrichment Score: 2.69			Count	P_Value	Benjamini
GOTERM_BP_ALL	immune system process	RT		8	1.1E-4	4.5E-1
GOTERM_BP_ALL	immune response	RT		6	8.5E-4	8.9E-1
GOTERM_BP_ALL	response to stimulus	RT		9	8.7E-2	1.0E0
Annotation Cluster 2	**Enrichment Score: 1.52**			**Count**	**P_Value**	**Benjamini**
GOTERM_BP_ALL	inflammatory response	RT		4	3.7E-3	9.8E-1
GOTERM_CC_ALL	extracellular region	RT		9	5.9E-3	9.9E-1
GOTERM_BP_ALL	response to wounding	RT		4	9.5E-3	1.0E0
GOTERM_BP_ALL	response to external stimulus	RT		4	3.1E-2	1.0E0
SP_PIR_KEYWORDS	Secreted	RT		5	5.6E-2	1.0E0
GOTERM_BP_ALL	defense response	RT		4	7.9E-2	1.0E0
GOTERM_BP_ALL	behavior	RT		3	8.0E-2	1.0E0
GOTERM_BP_ALL	response to stimulus	RT		9	8.7E-2	1.0E0
GOTERM_BP_ALL	response to stress	RT		4	1.1E-1	1.0E0

D16-Depleted in mesenchyme

Annotation Cluster 1	Enrichment Score: 14.88			Count	P_Value	Benjamini
GOTERM_CC_ALL	proteinaceous extracellular matrix	RT		54	1.5E-22	1.2E-19
GOTERM_CC_ALL	extracellular matrix	RT		54	5.5E-22	2.1E-19
SP_PIR_KEYWORDS	extracellular matrix	RT		37	6.2E-16	2.9E-13
GOTERM_CC_ALL	extracellular matrix part	RT		21	5.8E-10	9.2E-8
SP_PIR_KEYWORDS	Secreted	RT		82	1.4E-7	1.5E-5
Annotation Cluster 2	**Enrichment Score: 13.4**			**Count**	**P_Value**	**Benjamini**
GOTERM_BP_ALL	cell adhesion	RT		75	5.7E-19	9.9E-16
GOTERM_BP_ALL	biological adhesion	RT		75	5.7E-19	9.9E-16
SP_PIR_KEYWORDS	cell adhesion	RT		42	5.8E-11	1.3E-8
GOTERM_BP_ALL	cell-cell adhesion	RT		29	1.4E-7	2.1E-5

Table 3.4A. Transcription factors enriched or depleted in E14.5 stomach

Gene Symbol	Gene Title	S16-S14		D14-S14	
		FC	p-value	FC	p-value
Enriched					
Barx1	BarH-like homeobox 1	-2.09	0.002	-16.13	<0.001
Isl1	ISL1 transcription factor, LIM/homeodomain	-2.28	<0.001	-6.02	<0.001
Nr2f1	nuclear receptor subfamily 2, group F, member 1	-2.52	0.001	-4.00	<0.001
Trps1	Trichorhinophalangeal syndrome I (human)	-3.03	0.003	-3.25	<0.001
Mrg1	Myeloid ecotropic viral integration site-related gene 1	-6.00	<0.001	-3.04	<0.001
Sox5	SRY-box containing gene 5	-2.12	0.003	-2.72	0.001
Id4	inhibitor of DNA binding 4	-2.34	<0.001	-2.33	<0.001
Pbx1	pre B-cell leukemia transcription factor 1	-2.13	<0.001	-2.29	0.001
Klf12	Kruppel-like factor 12	-2.79	<0.001	-2.25	0.001
Id4	inhibitor of DNA binding 4	-2.58	<0.001	-2.16	<0.001
Scx	scleraxis	-2.22	<0.001	-2.05	0.002
Npas3	neuronal PAS domain protein 3	-2.38	<0.001	-2.04	<0.001
Zfp207	zinc finger protein 207	-4.91	<0.001	-2.01	0.028
Depleted					
Neurog3	neurogenin 3	2.17	<0.001	6.40	<0.001
Hod	homeobox only domain	6.53	<0.001	6.11	<0.001
Nr1h4	nuclear receptor subfamily 1, group H, member 4	4.60	<0.001	5.91	0.001
Cdx1	caudal type homeo box 1	2.07	<0.001	5.36	0.009
Hod	homeobox only domain	6.30	<0.001	4.70	<0.001
Ppargc1b	peroxisome proliferative activated receptor, gamma, coactivator 1 beta	3.45	<0.001	3.77	0.002
Cebpa	CCAAT/enhancer binding protein (C/EBP), alpha	2.03	<0.001	2.56	0.004

Table 3.4B. Transcription factors enriched or depleted in E14.5 duodenum

Gene Symbol	Gene Title	D16-D14		D14-S14		Epi-Mes	
		FC	p-value	FC	p-value	FC	p-value
Enriched							
Rfxdc1	Regulatory factor X domain containing 1	-3.25	<0.001	3.04	<0.001	2.49	0.075
Pitx2	paired-like homeodomain transcription factor 2	-2.89	<0.001	2.00	0.002	2.02	0.029
Ptf1a	pancreas specific transcription factor, 1a	-3.50	0.001	3.21	0.001	1.64	0.020
Neurog3	neurogenin 3	-2.47	0.015	6.40	<0.001	1.54	0.011
Hoxa4	homeo box A4	-4.78	<0.001	2.12	0.002	-5.12	<0.001
Cutl2	cut-like 2 (Drosophila)	-7.74	<0.001	2.05	<0.001	-6.46	<0.001
Hlx1	H2.0-like homeo box 1 (Drosophila)	-2.99	0.002	6.27	<0.001	-8.64	<0.001
Depleted							
Foxq1	Forkhead box Q1	2.66	0.002	-3.85	0.002	1.90	0.079
Nkx6-3	NK6 transcription factor related, locus 3 (Drosophila)	3.73	0.001	-3.02	<0.001	1.74	0.073
Prrx1	paired related homeobox 1	3.60	0.016	-2.30	0.003	1.51	0.072

Table 3.4C. Transcription factors enriched or depleted in E16.5 stomach

Gene Symbol	Gene Title	S16-S14		D16-S16	
		FC	p-value	FC	p-value
Enriched					
Foxq1	forkhead box Q1	4.07	<0.001	-4.88	<0.001
Ehf	ets homologous factor	2.84	0.002	-4.43	<0.001
Foxa3	forkhead box A3	2.38	0.001	-4.22	<0.001
Creb3l1	cAMP responsive element binding protein 3-like 1	2.36	<0.001	-4.09	<0.001
Bhlhb8	basic helix-loop-helix domain containing, class B, 8	4.91	<0.001	-4.05	<0.001
Stat5a	signal transducer and activator of transcription 5A	2.56	0.014	-3.89	0.003
Id1	inhibitor of DNA binding 1	2.64	0.001	-3.84	0.001
Nfix	nuclear factor I/X	2.42	0.007	-3.78	0.002
Tcfap2c	transcription factor AP-2, gamma	2.96	0.033	-3.68	0.017
Ovol2	ovo-like 2 (Drosophila)	2.57	0.001	-3.62	<0.001
Pir	pirin	2.07	0.001	-3.56	<0.001
Klf4	Kruppel-like factor 4 (gut)	4.14	0.001	-3.07	0.004
Ppard	peroxisome proliferator activator receptor delta	2.64	0.001	-2.95	<0.001
Nrip2	nuclear receptor interacting protein 2	3.78	0.003	-2.70	0.032
E2f5	E2F transcription factor 5	3.76	0.002	-2.58	0.013
Pparg	peroxisome proliferator activated receptor gamma	5.48	<0.001	-2.51	<0.001
Depleted					
Tbx3	T-box 3	-2.01	0.001	5.89	<0.001
Bach1	BTB and CNC homology 1	-3.10	0.004	3.33	0.004
Tle4	transducin-like enhancer of split 4, homolog of Drosophila E(spl)	-2.37	<0.001	2.03	0.001

Table 3.4D. Top 20 transcription factors enriched or depleted in E16.5 duodenum (part 1)

Gene Symbol	Gene Title	D16-D14		D16-S16		Epi-Mes	
		FC	p-value	FC	p-value	FC	p-value
Enriched							
Hnf4g	hepatocyte nuclear factor 4, gamma	3.66	0.002	13.19	<0.001	49.82	<0.001
Tcfec	transcription factor EC	5.45	0.003	41.65	<0.001	18.27	<0.001
Nr1h4	nuclear receptor subfamily 1, group H, member 4	9.7	<0.001	12.45	<0.001	13.04	<0.001
Creb3l3	cAMP responsive element binding protein 3-like 3	25.92	<0.001	18.95	<0.001	11.26	<0.001
Cdx1	caudal type homeo box 1	8.42	0.004	21.75	<0.001	9.56	<0.001
Plagl2	pleiomorphic adenoma gene-like 2	2.38	0.002	3.89	<0.001	7.06	0.001
Ppara	peroxisome proliferator activated receptor alpha	5.04	<0.001	9.09	<0.001	6.58	0.001
Nr1i3	nuclear receptor subfamily 1, group I, member 3	8.22	<0.001	8.46	<0.001	6.08	<0.001
Cebpa	CCAAT/enhancer binding protein (C/EBP), alpha	3.36	0.002	4.24	<0.001	5.96	<0.001
Nr2e3	nuclear receptor subfamily 2, group E, member 3	23.8	<0.001	18.68	<0.001	4.6	0.001
Mafb	v-maf musculoaponeurotic fibrosarcoma oncogene family, protein B	14.49	<0.001	14.49	<0.001	2.93	0.006
Solh	Small optic lobes homolog (Drosophila)	2.64	0.008	4.34	<0.001	2.84	0.005
Prrx1	paired related homeobox 1	9.57	<0.001	5.35	<0.001	2.73	0.007
Bach1	BTB and CNC homology 1	2.32	<0.001	3.84	<0.001	2.67	0.004
Maf	avian musculoaponeurotic fibrosarcoma (v-maf) AS42 oncogene homolog	5.38	<0.001	6.89	<0.001	2.41	0.001
Mlxipl	MLX interacting protein-like	2.93	0.004	7.87	<0.001	2.4	0.007
Esrra	estrogen related receptor, alpha	3.1	0.008	4.18	<0.001	2.39	0.037
Irf7	interferon regulatory factor 7	8.09	0.003	6.77	0.001	2.18	0.034
Nr1h3	nuclear receptor subfamily 1, group H, member 3	4.17	<0.001	4.45	<0.001	1.86	0.005
Ppargc1a	peroxisome proliferative activated receptor, gamma, coactivator 1 alpha	3.08	0.003	3.61	0.001	1.42	0.358

Table 3.4D. Top 20 transcription factors enriched or depleted in E16.5 duodenum (part 2)

Gene Symbol	Gene Title	D16-D14		D16-S16		Epi-Mes	
		FC	p-value	FC	p-value	FC	p-value
Depleted							
Foxa1	forkhead box A1	-4.04	<0.001	-8.35	<0.001	5.72	<0.001
Sox9	SRY-box containing gene 9	-3.42	<0.001	-4.77	<0.001	1.93	0.023
Isl1	ISL1 transcription factor, LIM/homeodomain	-2.7	<0.001	-8.86	<0.001	1.34	0.115
Elf5	E74-like factor 5	-2.13	0.012	-4.39	<0.001	1.29	0.07
Pitx1	paired-like homeodomain transcription factor 1	-3.07	0.007	-6.05	0.002	1.08	0.36
Foxa2	forkhead box A2	-3.42	<0.001	-8	<0.001	1.08	0.624
Barx1	BarH-like homeobox 1	-2.79	<0.001	-21.47	<0.001	-1.05	0.711
Nfe2l3	nuclear factor, erythroid derived 2, like 3	-2.91	0.002	-9.4	<0.001	-1.08	0.496
Mycl1	v-myc myelocytomatosis viral oncogene homolog 1, lung carcinoma derived	-3.67	<0.001	-3.87	<0.001	-1.39	0.01
Sox5	SRY-box containing gene 5	-3.29	0.006	-5.45	<0.001	-1.84	0.022
Sox2	SRY-box containing gene 2	-8.92	<0.001	-44.62	<0.001	-4.22	0.002
Ikzf2	IKAROS family zinc finger 2	-3.7	<0.001	-4.05	<0.001	-4.51	0.002
Tcf3	transcription factor 3	-6.03	<0.001	-5.35	<0.001	-4.74	0.001
Phox2a	paired-like homeobox 2a	-9.65	<0.001	-4.42	<0.001	-4.98	0.001
Ldb2	LIM domain binding 2	-7.96	<0.001	-3.83	<0.001	-6.36	0.005
Nr2f1	nuclear receptor subfamily 2, group F, member 1	-7.47	<0.001	-11.84	<0.001	-9.15	<0.001
Hoxc5	homeo box C5	-6.28	<0.001	-4.42	0.009	-9.46	<0.001
Trps1	trichorhinophalangeal syndrome I (human)	-7.09	<0.001	-4.55	<0.001	-11.3	<0.001
Nr2f2	nuclear receptor subfamily 2, group F, member 2	-3.97	<0.001	-6.84	<0.001	-15.9	<0.001
Nfix	nuclear factor I/X	-2.6	0.001	-5.77	<0.001	-18.5	<0.001

Table 3.5 Expression changes in genes for signaling factors during epithelial pyloric border formation

Gene Symbol	D16-D14		S16-S14		D14-S14		D16-S16		Epi-Mes		
	FC	p-value	FC	p-value	FC	p-value	FC	p-value	FC	p-value	Epi or Mes?
Fgf pathway											
Fgf1	4.49	0.00	1.32	0.02	-1.19	0.07	2.85	0.00	-1.51	0.12392	-
Fgf1	5.75	0.00	1.40	0.00	-1.05	0.56	3.91	0.00	1.21	0.15506	-
Fgf1	4.49	0.00	1.32	0.02	-1.19	0.07	2.85	0.00	-1.51	0.12392	-
Fgf13	-4.60	0.00	-1.86	0.00	1.13	0.05	-2.20	0.00	-47.59	0.00001	Mes
Fgf14	-1.15	0.01	2.16	0.00	-1.64	0.02	-4.10	0.00			-
Fgf14	-1.32	0.08	2.42	0.00	-2.03	0.00	-6.50	0.00			-
Fgfbp1	4.79	0.00	7.83	0.00	-2.67	0.02	-4.37	0.00			-
Fgfr1	-3.46	0.00	-1.35	0.13	-1.11	0.43	-2.84	0.00	-5.65	0.00001	Mes
Fgfr1	-3.05	0.01	-1.38	0.13	-1.11	0.39	-2.46	0.01	-35.52	0.00001	Mes
Fgfr1	-3.05	0.01	-1.38	0.13	-1.11	0.39	-2.46	0.01	-35.52	0.00001	Mes
Fgfr2	-3.75	0.00	-1.20	0.24	1.05	0.76	-2.98	0.00	-3.71	0.00020	Mes
Fgfr2	-3.16	0.00	-1.12	0.46	1.09	0.26	-2.59	0.00	-2.65	0.00003	Mes
Spred1	-2.49	0.00	-1.12	0.58	-1.12	0.54	-2.49	0.00	-6.43	0.00121	Mes
Spred1	-3.03	0.00	-2.17	0.00	-1.44	0.00	-2.01	0.00	-8.39	0.00067	Mes
Spred1	-3.14	0.01	-2.43	0.00	-1.74	0.02	-2.25	0.01	-8.52	0.00024	Mes
Spred1	-2.29	0.01	-1.19	0.38	-1.17	0.49	-2.25	0.00	-13.59	0.00125	Mes
Spred1	-2.28	0.00	-1.08	0.61	-1.31	0.16	-2.76	0.00	-6.18	0.00240	Mes
Spry1	-2.27	0.01	1.09	0.67	-1.14	0.05	-2.82	0.01	-5.04	0.00028	Mes

Gene Symbol	D16-D14		S16-S14		D14-S14		D16-S16		Epi-Mes		
	FC	p-value	FC	p-value	FC	p-value	FC	p-value	FC	p-value	Epi or Mes?
Notch pathway											
Acsl1	-12.31	0.00	-3.39	0.00	1.37	0.01	-2.66	0.01	-8.82	0.0001	Mes
Aph1c	2.71	0.02	1.04	0.36	1.13	0.16	2.94	0.02	1.68	0.017	-
Dlk1	-4.95	0.00	1.02	0.93	1.12	0.35	-4.49	0.00	-39.86	0.000	Mes
Dner	-2.52	0.00	-1.24	0.03	-1.03	0.67	-2.08	0.00	-2.33	0.024	Mes
Dner	-3.26	0.00	-1.22	0.02	-1.06	0.33	-2.84	0.01	-6.64	0.011	Mes
Dtx1	2.39	0.00	1.01	0.89	-1.13	0.35	2.08	0.00	1.16	0.373	-
Dtx3l	3.25	0.00	1.18	0.17	1.27	0.20	3.49	0.00	1.73	0.063	-
Dtx4	-2.03	0.00	-1.00	0.99	-1.09	0.01	-2.20	0.01	-2.14	0.013	Mes
Dtx4	-2.74	0.00	1.05	0.70	-1.08	0.06	-3.12	0.00	-4.30	0.001	Mes
Jag1	-1.51	0.01	-2.50	0.02	-2.04	0.01	-1.24	0.33	-6.08	0.007	Mes
Ncor2	-2.37	0.01	-1.01	0.95	1.05	0.65	-2.23	0.03	-3.15	0.001	Mes
Notch2	-4.18	0.00	-1.54	0.00	-1.12	0.08	-3.04	0.00	-6.21	0.000	Mes
Notch3	-3.74	0.00	-1.16	0.59	-1.18	0.35	-3.81	0.01	-4.57	0.003	Mes
Notch3	-2.34	0.01	-1.04	0.92	-1.22	0.28	-2.74	0.04	-4.81	0.002	Mes
Hedgehog pathway											
Shh	-8.94	0.00	1.30	0.01	1.75	0.00	-6.66	0.00	1.33	0.09306	-
Smo	-3.65	0.00	-1.61	0.00	1.03	0.72	-2.21	0.00	-4.86	0.00006	Mes
Gas1	-4.36	0.00	-1.67	0.00	1.11	0.27	-2.36	0.00	-19.45	0.00001	Mes
Boc	-4.86	0.00	-1.25	0.29	-1.05	0.72	-4.09	0.00	-5.18	0.00014	Mes
Gli1	-3.65	0.01	-1.14	0.43	1.04	0.83	-3.06	0.00	-6.31	0.00016	Mes
Gli2	-4.40	0.00	-1.99	0.01	1.04	0.72	-2.12	0.01	-3.44	0.00013	Mes
Gli3	-6.56	0.00	-2.70	0.00	-1.26	0.01	-3.07	0.00	-9.27	0.00002	Mes

Gene Symbol	D16-D14		S16-S14		D14-S14		D16-S16		Epi-Mes		
	FC	p-value	FC	p-value	FC	p-value	FC	p-value	FC	p-value	Epi or Mes?
Wnt pathway											
Fzd1	-3.15	0.00	-1.04	0.81	1.42	0.00	-2.14	0.00	-7.95	0.00013	Mes
Fzd2	-5.61	0.00	-2.00	0.00	-1.62	0.02	-4.55	0.00	-6.13	0.00021	Mes
Fzd2	-5.65	0.00	-1.95	0.00	-1.43	0.01	-4.16	0.00	-15.60	0.00007	Mes
Sfrp1	-5.45	0.00	-3.24	0.00	1.14	0.19	-2.17	0.00	-4.91	0.00076	Mes
Sfrp1	-8.13	0.00	-3.49	0.00	-1.62	0.02	-2.87	0.01	-6.13	0.00021	Mes
Sfrp1	-4.14	0.00	-2.20	0.00	-1.16	0.13	-2.18	0.00	-26.32	0.00001	Mes
Sfrp2	-7.70	0.00	-1.55	0.01	1.42	0.00	-6.37	0.00	-7.95	0.00013	Mes
Sfrp4	1.62	0.00	2.13	0.00	-1.91	0.01	-2.52	0.00			-
Sfrp5	-4.64	0.00	1.12	0.15	13.38	0.00	2.58	0.00	0.00	1.65050	-
Sfrp5	-4.71	0.00	-1.58	0.07	31.69	0.00	10.66	0.00	0.52	1.10068	-
Dkk2	-6.56	0.00	-2.05	0.00	1.06	0.37	-3.01	0.00	-10.04	0.00004	Mes
Dkk3	-2.57	0.00	1.11	0.49	1.24	0.03	-2.29	0.00	-8.05	0.00154	Mes
Dkk3	-2.01	0.00	1.33	0.10	1.14	0.19	-2.34	0.00	-4.91	0.00076	Mes
Tcf4	-3.81	0.00	-2.03	0.00	1.06	0.37	-2.29	0.01	-10.04	0.00004	Mes
Tcf4	-2.95	0.01	-1.95	0.00	1.24	0.03	-2.13	0.01	-8.05	0.00154	Mes
Tcf3	-6.03	0.00	-1.63	0.00	-1.43	0.01	-5.35	0.00	-15.60	0.00007	Mes
Grhl1	-1.49	0.01	2.73	0.01	-1.37	0.01	-5.57	0.00			-

Gene Symbol	D16-D14		S16-S14		D14-S14		D16-S16		Epi-Mes		
	FC	p-value	FC	p-value	FC	p-value	FC	p-value	FC	p-value	Epi or Mes?
Igf pathway											
Igf1	-3.24	0.00	-2.20	0.00	-2.91	0.00	-4.29	0.00	-5.56	0.00009	Mes
Igf1	-2.73	0.00	-2.14	0.00	-2.56	0.00	-3.27	0.00	-2.59	0.00022	Mes
Igf1r	-2.97	0.00	-1.39	0.03	-1.21	0.07	-2.60	0.00	-4.47	0.00005	Mes
Igf1r	-3.99	0.00	-2.28	0.00	-1.58	0.00	-2.76	0.00	-7.57	0.00002	Mes
Igfbp2	-5.07	0.00	-1.51	0.02	-2.33	0.00	-7.84	0.00	-22.44	0.00002	Mes
Igfbp4	-3.03	0.00	-1.28	0.03	-1.05	0.50	-2.49	0.00	-8.30	0.00003	Mes
Igfbp4	-4.06	0.00	-1.24	0.05	1.05	0.65	-3.12	0.00	-12.35	0.00004	Mes
Igfbp4	-2.37	0.00	-1.28	0.01	-1.15	0.05	-2.13	0.00	-18.14	0.00017	Mes
Igfbp4	-3.68	0.00	-1.44	0.01	-1.06	0.54	-2.71	0.00	-42.40	0.00009	Mes
Igfbp5	-4.32	0.00	-1.29	0.00	-1.94	0.00	-6.48	0.00	-40.42	0.00001	Mes
Igfbp6	1.69	0.04	2.09	0.01	-1.74	0.04	-2.15	0.00			-
Igfbp7	1.79	0.01	4.30	0.00	1.02	0.87	-2.34	0.00			-
Igfals	6.75	0.00	1.48	0.00	1.00	1.00	4.56	0.00	1.80	0.01116	-
Bmp pathway											
Bmp7	2.44	0.03	-1.40	0.03	-1.38	0.07	2.48	0.01	1.53	0.10845	-
Bmp7	3.77	0.00	-1.20	0.20	-1.43	0.03	3.16	0.00	4.73	0.00041	Epi
Bmpr1b	-2.01	0.00	-2.28	0.00	-5.83	0.00	-5.16	0.00	-1.24	0.02506	-
Id4	-3.17	0.00	-2.58	0.00	-2.16	0.00	-2.65	0.00	-20.86	0.00007	Mes
Id4	-3.41	0.00	-2.34	0.00	-2.33	0.00	-3.40	0.00	-33.37	0.00004	Mes
Twsg1	-2.25	0.02	-1.19	0.52	-1.24	0.16	-2.33	0.04	-3.10	0.00091	Mes
Twsg1	-2.29	0.00	-1.06	0.13	-1.40	0.00	-3.00	0.00	-4.07	0.00044	Mes
Twsg1	-3.37	0.01	-1.21	0.29	-1.41	0.04	-3.93	0.00	-8.57	0.00165	Mes

Table 3.6A. Genes enriched or depleted in E14.5 border

Gene Symbol	Gene Title	B14-S14		B14-D14	
		FC	p-value	FC	p-value
Enriched					
Gata3	GATA binding protein 3	6.75	0.007	10.35	0.004
Nkx2-5	NK2 transcription factor related, locus 5 (Drosophila)	5.64	0.001	6.58	0.001
Lgals6	lectin, galactose binding, soluble 6	7.12	0.001	4.57	0.001
Lect1	leukocyte cell derived chemotaxin 1	4.27	0.011	4.45	0.009
Skiv2l2	superkiller viralicidic activity 2-like 2 (S. cerevisiae)	3.88	0.019	3.84	0.013
2210407C18Rik	RIKEN cDNA 2210407C18 gene	5.01	0.002	2.87	0.008
Usp3	Ubiquitin specific peptidase 3	2.23	0.003	2.86	0.003
Arrdc3	Arrestin domain containing 3	2.57	0.006	2.84	0.006
9030612M13Rik	RIKEN cDNA 9030612M13 gene	2.19	0.017	2.45	0.024
EG665081	predicted gene, EG665081	2.19	0.004	2.36	0.013
Grem1	gremlin 1	7.13	0.002	2.26	0.037
AK135583	RIKEN full-length enriched library, clone:7030417P12	2.05	0.016	2.08	0.010
Skap2	src family associated phosphoprotein 2	2.56	0.005	2.06	0.013
Depleted					
Mreg	melanoregulin	-3.33	0.000	-3.12	0.000

Highlighted genes were examined by in situ hybridization, immunostaining or using a LacZ-tagged allele

Table 3.6B Genes enriched or depleted in E16.5 border

Gene Symbol	Gene Title	B16-S16		B16-D16	
		FC	p-value	FC	p-value
Enriched					
Nepn	nephrocan	17.06	0.000	16.73	0.000
1190003M12Rik	RIKEN cDNA 1190003M12 gene	2.48	0.002	12.56	0.002
Col9a1	procollagen, type IX, alpha 1	4.63	0.000	10.41	0.001
Gdf10	growth differentiation factor 10	2.74	0.001	8.62	0.000
1810065E05Rik	RIKEN cDNA 1810065E05 gene	4.51	0.000	6.56	0.000
LOC630963	similar to spectrin alpha 1	5.80	0.000	5.54	0.001
Crabp1	cellular retinoic acid binding protein I	2.14	0.013	5.53	0.003
Slc4a1	solute carrier family 4 (anion exchanger), member 1	4.11	0.033	5.28	0.028
Gypa	glycophorin A	4.00	0.013	4.59	0.016
E130306M17Rik	RIKEN cDNA E130306M17 gene	2.80	0.003	4.29	0.008
Upk1a	uroplakin 1A	4.91	0.000	4.09	0.000
Gm784	gene model 784, (NCBI)	2.24	0.001	4.08	0.000
Malat1	Metastasis associated lung adenocarcinoma transcript 1 (non-coding RNA)	3.23	0.023	4.01	0.018
Moxd1	monooxygenase, DBH-like 1	2.30	0.003	3.95	0.001
A730054J21Rik	RIKEN cDNA A730054J21 gene	2.47	0.000	3.95	0.002
Snca	synuclein, alpha	2.40	0.015	3.87	0.007
Hemgn	hemogen	3.14	0.026	3.79	0.023
Hnt	neurotrimin	2.32	0.000	3.77	0.000
Rprm	reprimo, TP53 dependent G2 arrest mediator candidate	2.14	0.001	3.67	0.002
Pln	phospholamban	2.00	0.001	3.63	0.000
Hbb-b1	hemoglobin, beta adult major chain	2.44	0.015	3.60	0.013
Eraf	erythroid associated factor	2.88	0.033	3.60	0.020
Ddx6	DEAD (Asp-Glu-Ala-Asp) box polypeptide 6	2.68	0.016	3.54	0.014

Gene Symbol	Gene Title	B16-S16		B16-D16	
		FC	p-value	FC	p-value
Cenpe	centromere protein E	2.25	0.007	3.42	0.009
Alas2	aminolevulinic acid synthase 2, erythroid	2.89	0.012	3.41	0.011
Gata3	GATA binding protein 3	2.55	0.006	3.39	0.002
Ror1	Receptor tyrosine kinase-like orphan receptor 1	2.29	0.004	3.33	0.014
Kctd12b	potassium channel tetramerisation domain containing 12b	2.49	0.026	3.28	0.005
BC056349	cDNA sequence BC056349	2.25	0.001	3.26	0.004
Spred2	sprouty-related, EVH1 domain containing 2	2.87	0.003	3.22	0.003
Msi2	Musashi homolog 2 (Drosophila)	2.25	0.005	3.03	0.007
Neto2	neuropilin (NRP) and tolloid (TLL)-like 2	2.24	0.004	2.98	0.005
Prkg1	protein kinase, cGMP-dependent, type I	2.33	0.001	2.97	0.004
Rgs13	regulator of G-protein signaling 13	3.50	0.005	2.90	0.036
Bhmt	betaine-homocysteine methyltransferase	2.56	0.030	2.89	0.036
Rbm5	RNA binding motif protein 5	3.15	0.033	2.88	0.017
Pap	pancreatitis-associated protein	63.15	0.000	2.80	0.002
Nkx2-5	NK2 transcription factor related, locus 5 (Drosophila)	2.49	0.000	2.76	0.000
Hba-x	hemoglobin X, alpha-like embryonic chain in Hba complex	2.17	0.033	2.70	0.016
Pnliprp2	pancreatic lipase-related protein 2	21.36	0.000	2.70	0.016
Grem1	gremlin 1	4.90	0.000	2.69	0.038
Cldn9	claudin 9	2.53	0.001	2.67	0.008
Actc1	actin, alpha, cardiac	2.04	0.011	2.58	0.011
Cutl2	cut-like 2 (Drosophila)	2.09	0.000	2.44	0.001
Rbbp4	retinoblastoma binding protein 4	2.56	0.046	2.44	0.012
Uck2	Uridine-cytidine kinase 2	2.19	0.027	2.42	0.027
Ambp	alpha 1 microglobulin/bikunin	8.72	0.000	2.39	0.002
Wfdc15	WAP four-disulfide core domain 15	2.90	0.000	2.34	0.001
Rhag	Rhesus blood group-associated A glycoprotein	2.12	0.044	2.33	0.047

Gene Symbol	Gene Title	B16-S16		B16-D16	
		FC	p-value	FC	p-value
Cftr	cystic fibrosis transmembrane conductance regulator homolog	3.73	0.002	2.33	0.046
Unc5c	Unc-5 homolog C (C. elegans)	2.48	0.020	2.28	0.016
Bat2d	BAT2 domain containing 1	2.34	0.025	2.28	0.047
Nr2c2	nuclear receptor subfamily 2, group C, member 2	2.02	0.017	2.25	0.001
Gsta3	glutathione S-transferase, alpha 3	2.97	0.004	2.24	0.027
2610203C20Rik	RIKEN cDNA 2610203C20 gene	2.40	0.006	2.20	0.007
Chd4	chromodomain helicase DNA binding protein 4	2.46	0.039	2.19	0.037
3100002J23Rik	RIKEN cDNA 3100002J23 gene	2.65	0.001	2.19	0.009
4931406P16Rik	RIKEN cDNA 4931406P16 gene	2.92	0.005	2.18	0.003
Mapre2	microtubule-associated protein, RP/EB family, member 2	2.13	0.009	2.14	0.008
Prpf19	PRP19/PSO4 pre-mRNA processing factor 19 homolog (S. cerevisiae)	2.02	0.023	2.12	0.030
AW822216	Expressed sequence AW822216	3.37	0.001	2.10	0.021
Epha7	Eph receptor A7	2.14	0.002	2.10	0.005
BC057627	cDNA sequence BC057627	2.14	0.026	2.10	0.009
Spna1	spectrin alpha 1 /// similar to spectrin alpha 1	2.73	0.003	2.09	0.021
Nkx6-3	NK6 transcription factor related, locus 3 (Drosophila)	2.00	0.009	2.08	0.001
Ildr1	immunoglobulin-like domain containing receptor 1	2.85	0.002	2.07	0.023
Slc12a8	solute carrier family 12 (potassium/chloride transporters), member 8	2.77	0.000	2.04	0.035
Pten	phosphatase and tensin homolog	3.36	0.010	2.02	0.002
Hnrpab	heterogeneous nuclear ribonucleoprotein A/B	2.67	0.030	2.00	0.014
Depleted					
Pdia2	protein disulfide isomerase associated 2	-5.69	0.001	-2.88	0.030
Npnt	nephronectin	-2.85	0.032	-2.05	0.050

Bibliography

Allen BL, Tenzen T, McMahon AP. 2007. The Hedgehog-binding proteins Gas1 and Cdo cooperate to positively regulate Shh signaling during mouse development. *Genes Dev* 21: 1244-57

Bai CB, Auerbach W, Lee JS, Stephen D, Joyner AL. 2002. Gli2, but not Gli1, is required for initial Shh signaling and ectopic activation of the Shh pathway. *Development* 129: 4753-61

Boesze-Battaglia K, Song H, Sokolov M, Lillo C, Pankoski-Walker L, et al. 2007. The tetraspanin protein peripherin-2 forms a complex with melanoregulin, a putative membrane fusion regulator. *Biochemistry* 46: 1256-72

Braunstein EM, Qiao XT, Madison B, Pinson K, Dunbar L, Gumucio DL. 2002. Villin: A marker for development of the epithelial pyloric border. *Dev Dyn* 224: 90-102

Canning CA, Lee L, Irving C, Mason I, Jones CM. 2007. Sustained interactive Wnt and FGF signaling is required to maintain isthmic identity. *Dev Biol* 305: 276-86

Chi NC, Shaw RM, De Val S, Kang G, Jan LY, et al. 2008. Foxn4 directly regulates tbx2b expression and atrioventricular canal formation. *Genes Dev* 22: 734-9

Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, et al. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4: P3

Goodrich LV, Milenkovic L, Higgins KM, Scott MP. 1997. Altered neural cell fates and medulloblastoma in mouse patched mutants. *Science* 277: 1109-13

Harada T, Chelala C, Bhakta V, Chaplin T, Caulee K, et al. 2008. Genome-wide DNA copy number analysis in pancreatic cancer using high-density single nucleotide polymorphism arrays. *Oncogene* 27: 1951-60

Hiraki Y, Shukunami C. 2005. Angiogenesis inhibitors localized in hypovascular mesenchymal tissues: chondromodulin-I and tenomodulin. *Connect Tissue Res* 46: 3-11

Ho IC, Pai SY. 2007. GATA-3 - not just for Th2 cells anymore. *Cell Mol Immunol* 4: 15-29

Huang da W, Sherman BT, Tan Q, Kir J, Liu D, et al. 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res* 35: W169-75

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4: 249-64

Ishii Y, Rex M, Scotting PJ, Yasugi S. 1998. Region-specific expression of chicken Sox2 in the developing gut and lung epithelium: regulation by epithelial-mesenchymal interactions. *Dev Dyn* 213: 464-75

Joyner AL, Liu A, Millet S. 2000. Otx2, Gbx2 and Fgf8 interact to position and maintain a mid-hindbrain organizer. *Curr Opin Cell Biol* 12: 736-41

Kawazoe Y, Sekimoto T, Araki M, Takagi K, Araki K, Yamamura K. 2002. Region-specific gastrointestinal Hox code during murine embryonal gut development. *Dev Growth Differ* 44: 77-84

Kim BM, Buchner G, Miletich I, Sharpe PT, Shivdasani RA. 2005. The stomach mesenchymal transcription factor Barx1 specifies gastric epithelial identity through inhibition of transient Wnt signaling. *Dev Cell* 8: 611-22

Kim BM, Mao J, Taketo MM, Shivdasani RA. 2007a. Phases of canonical Wnt signaling during the development of mouse intestinal epithelium. *Gastroenterology* 133: 529-38

Kim BM, Miletich I, Mao J, McMahon AP, Sharpe PA, Shivdasani RA. 2007b. Independent functions and mechanisms for homeobox gene Barx1 in patterning mouse stomach and spleen. *Development* 134: 3603-13

Lees C, Zacharias W, Tremelling M, Noble CL, Nimmo ER, et al. 2008. Analysis of germline GLI1 variation implicates Hedgehog signalling in the regulation of intestinal inflammatory pathways. *PLOS Med* In press

Li X, Madison BB, Zacharias W, Kolterud A, States D, Gumucio DL. 2007. Deconvoluting the intestine: molecular evidence for a major role of the mesenchyme in the modulation of signaling cross talk. *Physiol Genomics* 29: 290-301

Lincoln J, Florer JB, Deutsch GH, Wenstrup RJ, Yutzey KE. 2006. ColVa1 and ColXIa1 are required for myocardial morphogenesis and heart valve development. *Dev Dyn* 235: 3295-305

Luebke-Wheeler J, Zhang K, Battle M, Si-Tayeb K, Garrison W, et al. 2008. Hepatocyte nuclear factor 4alpha is implicated in endoplasmic reticulum stress-induced acute phase response by regulating expression of cyclic adenosine monophosphate responsive element binding protein H. *Hepatology* 48: 1242-50

Madison BB, Braunstein K, Kuizon E, Portman K, Qiao XT, Gumucio DL. 2005. Epithelial hedgehog signals pattern the intestinal crypt-villus axis. *Development* 132: 279-89

Madison BB, Dunbar L, Qiao XT, Braunstein K, Braunstein E, Gumucio DL. 2002. Cis elements of the villin gene control expression in restricted domains of the vertical (crypt) and horizontal (duodenum, cecum) axes of the intestine. *J Biol Chem* 277: 33275-83

Mandalaywala NV, Chang S, Snyder RG, Levendusky MC, Voigt JM, Dearborn RE, Jr. 2008. The tumor suppressor, vitamin D3 up-regulated protein 1 (VDUP1), functions downstream of REPO during Drosophila gliogenesis. *Dev Biol* 315: 489-504

Mochida Y, Parisuthiman D, Kaku M, Hanai J, Sukhatme VP, Yamauchi M. 2006. Nephrocan, a novel member of the small leucine-rich repeat protein family, is an inhibitor of transforming growth factor-beta signaling. *J Biol Chem* 281: 36044-51

Moniot B, Biau S, Faure S, Nielsen CM, Berta P, et al. 2004. SOX9 specifies the pyloric sphincter epithelium through mesenchymal-epithelial signals. *Development* 131: 3795-804

Nechushtan H, Razin E. 2002. The function of MITF and associated proteins in mast cells. *Mol Immunol* 38: 1177-80

Nicassio F, Corrado N, Vissers JH, Areces LB, Bergink S, et al. 2007. Human USP3 is a chromatin modifier required for S phase progression and genome stability. *Curr Biol* 17: 1972-7

Park HL, Bai C, Platt KA, Matise MP, Beeghly A, et al. 2000. Mouse Gli1 mutants are viable but have defects in SHH signaling in combination with a Gli2 mutation. *Development* 127: 1593-605

Partington GA, Fuller K, Chambers TJ, Pondel M. 2004. Mitf-PU.1 interactions with the tartrate-resistant acid phosphatase gene promoter during osteoclast differentiation. *Bone* 34: 237-45

Parviz F, Matullo C, Garrison WD, Savatski L, Adamson JW, et al. 2003. Hepatocyte nuclear factor 4alpha controls the development of a hepatic epithelium and liver morphogenesis. *Nat Genet* 34: 292-6

Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, et al. 2006. TM4 microarray software suite. *Methods Enzymol* 411: 134-93

Saeed AI, Sharov V, White J, Li J, Liang W, et al. 2003. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34: 374-8

Scholpp S, Lohs C, Brand M. 2003. Engrailed and Fgf8 act synergistically to maintain the boundary between diencephalon and mesencephalon. *Development* 130: 4881-93

Silberg DG, Swain GP, Suh ER, Traber PG. 2000. Cdx1 and cdx2 expression during intestinal development. *Gastroenterology* 119: 961-71

Smith DM, Nielsen C, Tabin CJ, Roberts DJ. 2000. Roles of BMP signaling and Nkx2.5 in patterning at the chick midgut-foregut boundary. *Development* 127: 3671-81

Stehberger PA, Shmukler BE, Stuart-Tilley AK, Peters LL, Alper SL, Wagner CA. 2007. Distal renal tubular acidosis in mice lacking the AE1 (band3) Cl-/HCO3- exchanger (slc4a1). *J Am Soc Nephrol* 18: 1408-18

Tachibana M. 2000. MITF: a stream flowing for pigment cells. *Pigment Cell Res* 13: 230-40

Theodosiou NA, Tabin CJ. 2005. Sox9 and Nkx2.5 determine the pyloric sphincter epithelium under the control of BMP signaling. *Dev Biol* 279: 481-90

Trokovic R, Jukkola T, Saarimaki J, Peltopuro P, Naserke T, et al. 2005. Fgfr1-dependent boundary cells between developing mid- and hindbrain. *Dev Biol* 278: 428-39

Veranic P, Romih R, Jezernik K. 2004. What determines differentiation of urothelial umbrella cells? *Eur J Cell Biol* 83: 27-34

Wang BE, Wang XD, Ernst JA, Polakis P, Gao WQ. 2008. Regulation of epithelial branching morphogenesis and cancer cell growth of the prostate by Wnt signaling. *PLoS ONE* 3: e2186

Wells JM, Melton DA. 1999. Vertebrate endoderm development. *Annu Rev Cell Dev Biol* 15: 393-410

Yan D, Wiesmann M, Rohan M, Chan V, Jefferson AB, et al. 2001. Elevated expression of axin2 and hnkd mRNA provides evidence that Wnt/beta -catenin signaling is activated in human colon tumors. *Proc Natl Acad Sci U S A* 98: 14973-8

Yang CT, Hindes AE, Hultman KA, Johnson SL. 2007. Mutations in gfpt1 and skiv2l2 cause distinct stage-specific defects in larval melanocyte regeneration in zebrafish. *PLoS Genet* 3: e88

Zhang J, Rosenthal A, de Sauvage FJ, Shivdasani RA. 2001. Downregulation of Hedgehog signaling is required for organogenesis of the small intestine in Xenopus. *Dev Biol* 229: 188-202

Zhang K, Shen X, Wu J, Sakaki K, Saunders T, et al. 2006. Endoplasmic reticulum stress activates cleavage of CREBH to induce a systemic inflammatory response. *Cell* 124: 587-99

CHAPTER IV

CONCLUSION AND DISCUSSION

The work presented in this thesis has used microarray technology and bioinformatics techniques to study gut development using the mouse model. Development is a complex process and involves constant changes in gene expression. In the context of a tissue or organ, each cell is undergoing its own developmental program and each cell is influencing the neighboring cells around it. Thus, each cell takes in and gives out signals. With time, each cell changes in specific ways in response to those signals. Since any organ is made up of many different types of cells that all are changing in a dynamic way, this presents a difficult puzzle for the developmental biologist to unravel. We have tried to reduce this complexity in two ways: a) by separating major tissues (e.g., epithelium from mesenchyme) and b) by restricting our analysis to a major developmental event occurring at a specific space and time (e.g., epithelial pyloric border formation). Both approaches gave new insight into gut development and generated new questions for future study. Both studies also highlighted the strength as well as the shortcomings of bioinformatic approaches to understanding organ development.

In our first study, we categorized gene expression patterns in epithelium vs. mesenchyme. We know that both of these two tissues use soluble signals to talk to each other. Thus, understanding which signals come from which tissue and clarifying which tissue can respond to which signal is crucial for understanding and

further investigating the instructional crosstalk between these two tissues. We found that the mesenchyme expresses most of the regulators or inhibitors of many signaling pathways. In addition, we found that the mesenchyme is very generic tissue from a gene expression standpoint, while the epithelium is highly organ-specific. A challenge now is to understand how such a generic tissue can play such a powerful role in patterning such a specific one. There are at least three possible answers to this question. First, it is possible that though intestinal mesenchyme may express a very similar set of genes as for example stomach mesenchyme, perhaps the expression ratios of specific inhibitors to the corresponding signals are different in stomach and intestinal mesenchyme (similar expression level of signals but different level of inhibitors) and this difference could be instructive. Another possibility is that mesenchymal instructional control is mediated by a certain gene but perhaps it is expressed at low levels (perhaps due to expression in only a specific type of cells). We certainly did not examine all mesenchymal genes in the mouse genome; some genes are not on the chip; and some are not yet characterized (Riken genes, etc). It is even possible that the most important mesenchymal regulator is present in our list of candidate genes but we did not recognize it as important because it is not yet functionally characterized. Of course, instructional control is probably gained by the integrated use of multiple pathways, rather than one "intestine specific" signal. Finally, it is possible that the epithelium directs its own specificity by talking to the mesenchyme. We know that the epithelium does indeed have instructive power at some particular stages. Perhaps the epithelial signal at these stages acts to "blueprint" the surrounding mesenchyme so that mesenchyme expresses a specific combination of soluble signaling factors and/or inhibitors that then act back on the epithelium. For example, our microarray shows that the epithelium expresses both Ihh and Shh as well

as Bmp7. Any or all of these could pattern the mesenchyme to be “intestinal mesenchyme”.

Experimentally and computationally, there are several ways to approach these important questions. First, we have developed methods to culture separated mesenchyme and epithelium (Madison, et al., 2005). This makes it possible to test candidate epithelial signals on the cultured mesenchyme or vice versa. For example, important signals from the epithelium, such as Hh or Bmp7, can be added to isolated cultured mesenchyme. Then gene expression changes (or time series analysis) can be done using classical QRT-PCR or microarray technologies. Furthermore, it would be possible to test whether the conditioned mesenchyme could “instruct” stomach epithelium to become intestinal (that is, to express intestinal genes). This could be also combined with strategies to alter either the mesenchyme or the epithelium by using tissue from mouse knockouts or by using siRNA knockdowns.

Second, the whole embryonic small intestine can be cultured and agarose beads soaked with soluble signaling proteins or inhibitors can be placed on the intact intestine. The signaling protein will then diffuse from the beads into the mesenchymal tissue. If it is an important instructional signal, it will affect the development of the small intestine. This is a good way to test if a specific signal or inhibitor is critical but it cannot be applied easily to combinations of inhibitors or signals. Recent work in our lab has used this approach to study the ability of mesenchymal cells to direct the formation of epithelial villi via the Bmp pathway. This work is not published yet, but it is exciting because it is uncovering new information about villus formation and several of the candidate genes being tested are derived from the epithelium mesenchyme array studies described in Chapter Two of this thesis.

Another important outcome of the epithelial/mesenchymal compartmentalization study is that we identified transcription factors that are likely to be important in intestinal epithelial development. The role of Hnf4α had already been well studied in the liver (Parviz et al 2003) and colon (Garrison et al 2006). The knockout of this factor in these tissues leads to failure to differentiate fully. The knockout has no small intestinal phenotype, probably because the intestine has Hnf4γ as well, and these two factors are known to bind to the same site. However, from the importance in these other tissues, and from the very high level of expression of these Hnf4 factors in epithelium, we can guess that they play critical roles in intestinal differentiation as well. The fact that we find many intestinal epithelial genes with binding sites for Hnf4 in their proximal promoters is more evidence that they are functionally important. From work in the Kaufman laboratory here at the University of Michigan, we now know that at least one of the factors that Hnf4 regulates is Creb3l3, a factor that is also highly enriched in the intestinal epithelium (Luebke-Wheeler et al 2008).

It will be important to further determine which intestinal genes are regulated by Hnf4α and Hnf4γ. We have the Hnf4a knockout mice in the lab. Though they have no phenotype, it might be useful to use these mice and their wild type littermates to compare Hnf4α targets in epithelium using ChIP-on-Chip arrays. The knockout would provide a useful control. It also could be interesting to ChIP with Hnf4γ, as it might be that in the absence of Hnf4α, the Hnf4γ will bind additional targets.

There are, of course, many other potentially interesting transcription factors to analyze functionally. We have good tools in the lab for overexpressing genes or knocking out genes in the intestinal epithelium. The villin gene fragment which we characterized (Madison et al 2002) is now widely used in the studies of intestinal

development since it drives high level expression of genes in all cells of the intestinal epithelium (crypt to villus tip) in both large and small intestine, but not stomach. We have also developed a villin-Cre transgenic line that is based on this promoter and allows us to delete any gene in an intestine specific way. This specificity will be important because many of the transcription factors that we identified are known to cause embryonic lethality when they are generally deleted.

The gene expression profiles in epithelium and mesenchyme presented in Chapter 2 using microarray technology demonstrated the dramatic difference between these two tissues and offered valuable information for understanding the crosstalk between them. However, expression for most genes in embryonic development changes dynamically and our microarray data only examined one time point (E18.5) in late development. One of the interesting applications of microarray is definition of genetic pathways by combining precise temporal gene profiling and bioinformatics analysis. A time series study can provide information about the early and late expression profiles that may show activation and repression of transcription, suggesting that early-expressing transcription factors regulate the late-expression of other genes(Cho et al 1998). It would therefore be informative if we carry out a time series analysis for expression profiles of separated epithelium and mesenchyme. We know that villi begin to form at E14.5; the epithelial pyloric border forms about E16.5 (Braunstein et al 2002); and crypts form in the first week after birth. It seems that E14.5, E15.5, E16.5, P0 (birth) and P7 (one week after birth) would be critical time points for mouse gut development. Therefore, we can separate the epithelium and mesenchyme at these time points and perform microarray experiments on these points. The time series data from this microarray study will provide dynamic expression profiles and will be helpful to further investigate the back-and-forth crosstalk between these two tissues during these

important developmental events and understand the functions of genes based on the dynamic expression changes. Another strong advantage of a time series study is that clustering is more powerful and sensitive for identifying co-expressed or co-regulated genes (Cho et al 1998). After clustering analysis on this time series experiment, the resultant clusters can be analyzed by performing GO term or pathway analysis for functional similarity or relevance. If the functions and roles for genes in one cluster are highly similar or relevant for some biological function, transcription factor binding sites (TFBS) analysis can be applied given that the binding sequences for some relevant transcription factors are known. If we do not know the binding sequence or potential transcription factors, searching the common conserved sequences in promoter regions in the same group of genes will be another option. Through these analyses, important transcription factors and regulation mechanisms could be revealed. However, this is not always so easy. We did spend considerable time looking for possible functional connections for specific transcription factors using Genomatix tools and were not successful in many cases. This is discussed in more detail below.

Another consideration is that we used the entire small intestine for our analysis of the epithelium/mesenchyme separation array. But it is well known that there are regional differences in gene expression in duodenum, jejunum and ileum. In addition, during development, the more anterior regions differentiate first, with the more posterior regions lagging a day or two behind. By taking the whole intestine, we are not getting a “pure” but an average picture of the developmental steps. Even though it would be difficult and time consuming, we should consider restricting future analysis to just one segment (e.g. the duodenum). The information lost by not taking the entire intestine is probably less important than the resolution that is gained by studying just a

small region at multiple times. We may include different regions of intestine to recover the lost information which will reveal the expression difference of proximal and distal regions of intestine.

We saw how valuable restricting the spatial region was in our study of the pyloric border. In that study, detailed in Chapter 3, we determined global transcriptional changes that occur during an important patterning step in gut development: the formation of the epithelial pyloric border. In this analysis, we were able to make use of the epithelial/mesenchymal categorization to further understand the dramatic gene expression changes that took place in the duodenum during epithelial pyloric border formation. As in the first study, we learned many new details about gene expression in the developing gut, but in many ways, this study was more exciting because very little is known about how this border is formed and our study allowed us to generate several new hypotheses for future functional testing.

First, we discovered that the formation of the pyloric border occurs concomitantly with the up-regulation of about 2000 genes in the duodenal epithelium. Since many of these genes encode molecules that function in intestine-specific processes of absorption and metabolism, it appears that this major transcriptional event was essentially a gain of intestinal identity. We uncovered several novel transcription factors that might play a role in this process, which we call “intestinalization”.

One of these transcription factors, Tcfec, is expressed primarily in E16.5 duodenal epithelium. Currently its role in intestinal development is unknown, but its family members are known to be important in giving identity to other tissues such as retinal pigment cells (Tachibana 2000), mast cells (Nechushtan & Razin 2002) and osteoclasts (Partington et al 2004). The knockout of this gene was not informative (Steingrimsson et al 2002), probably because these factors are known to

heterodimerize and other family members are present in the intestine, though none are expressed as highly as Tcfec is. It would be of interest to perform a more detailed time series analysis of this process of intestinalization; we may identify a cluster of genes that get upregulated shortly after Tcfec does. Another approach is to transfect Tcfec into an intestinal cell line in culture and to examine the genes that are up-regulated by Tcfec. The CaCo2 line would be interesting for this, since published array studies indicate that this line does differentiate in culture and during differentiation; these cells express many of the same genes that we see in our array, including Tcfec (Fleet et al 2003, Saaf et al 2007). It would also be important to find out if a dominant negative form of Tcfec could inhibit differentiation either in these cells or when introduced into mice using the villin promoter. Once we are sure which genes are regulated by Tcfec, the actual cis targets can be identified by a combination of ChIP, sequence analysis, evolutionary conservation and gel shift studies.

Similar studies could be carried out for several of the other transcription factors that we identified in the duodenum. Creb3l3, for example, is a gene that seems to be involed in an ER stress pathway (Zhang et al 2006). It has not previously been considered that such a robust process like intestinalization, which involves the upregulation of over a thousand genes, might require that the endoplasmic reticulum change its state to become a large scale factory for proteins. It will be interesting to determine whether removing Creb3l3 prevents effective intestinalization.

One of the interesting things about the stomach-border-intestine microarray experiments is that dramatic changes of gene expression happened in intestine while only a small number of genes changed in stomach and border. This may imply that the stomach epithelium might be actively repressed from forming into intestine. As we discuss in Chapter 3, this might have implications for intestinal metaplasia, a

pathological condition in which the stomach expresses intestinal genes. These metaplasias are precursors to cancer, so it is important to determine how they initiate. We hope that some of the interesting genes that we have identified in this study could provide clues. For example, we might look for proteins that are expressed in the stomach epithelium and that can act as repressors. We did not follow up very much on the stomach genes determined here, since most of the past work of our lab has been about intestinal development. A first step to study stomach development further would be to perform the epithelial/mesenchymal separation that we did for duodenum. Then we could identify specific transcription factors that are expressed in stomach epithelium, determine which of these are known to be repressors and begin testing these factors functionally. An interesting candidate in this regard is Sox21. The Sox factors can be activators or repressors, and Sox21 seems to be very high in the stomach, but not expressed in duodenum. Perhaps the overexpression of Sox21 in CaCo2 cells could prevent their differentiation into intestine. Or maybe the deletion of Sox21 would allow the stomach to differentiate into intestine.

Another interesting finding from our border microarray was the identification of some signaling pathways that might be important in border formation or in intestinalization. Two recent studies in the literature from the Shivdasani group suggest that canonical Wnt signals are important in both stomach (Kim et al 2005a) and intestinal (Kim et al 2007a) development. In the stomach, transient expression of Barx1 at E12.5 inhibits Wnt signaling by upregulating Sfrp1 and Sfrp2. This reduction in Wnt signals is required for the differentiation of the gastric epithelium (Kim et al 2005a). Of course, this event occurs well before pyloric border formation at E16.5 and the Kim et al. study did not examine Wnt signaling at later time points. In a different study of the intestine, this same group of investigators put forward an unconventional idea: that

just after development of the intestinal villi, proliferation in the intervillus region does NOT depend on canonical Wnt signals (Kim et al 2007a). Using TOPGAL and Axin2-LacZ reporter mice (that express β-galactosidase under control of Tcf binding sites or Axin2 promoters, respectively), they provide evidence that the tips of the developing villi transduce canonical Wnt signals, but the intervillus region does not (Kim et al 2007a). This is the opposite of what has been conventionally accepted (van de Wetering et al 2002). According to their data, this is true from E16.5 until postnatal day three (P3). Interestingly, in this entire study, the authors never use in situ hybridization to look for evidence of active canonical Wnt signaling. They only rely on the reporter mice and on immunohistochemical stainings, both which can be problematic. Investigators in our laboratory have seen that precisely at E16.5, there is a very large increase in background X-gal staining in the intestinal epithelium. It is likely that one of the many intestinal epithelial genes that are up-regulated at this time is a gene with endogenous β-gal activity.

Our in situ analysis of Axin2 expression, shown in Chapter 3, causes us to seriously doubt the conclusions of the Kim et al. study on Wnt signaling in intestine (Kim et al 2007a). Axin2 is a known target gene of Wnt signaling (Jho et al 2002, Lustig et al 2002, Yan et al 2001) and is commonly thought to be the best and most specific indicator of canonical Wnt pathway activity (Wang et al 2008). When we first looked at Axin2 expression in the microarray, we found that it is expressed at very low levels if at all at E14.5. In fact, the in situ hybridization bears this out. At E16.5, we could detect a bit more Axin2 expression in the array output, but we still did not see a difference in expression level between stomach and intestine. However, several other Wnt modulators (especially the Sfrps, which are soluble Wnt inhibitors) did seem to change between stomach and duoduenum. Therefore, we decided to examine Axin2

expression at E16.5 by in situ hybridization to determine whether there is a difference in Wnt signaling from stomach to intestine. We found no evidence for changes in Wnt signaling across the border, but our results provided important clarification on the question of Wnt signaling on villus tips vs. intervillus regions: Axin2 expression (and therefore, canonical Wnt pathway activity) is found exclusively in the intervillus region as originally thought, not the villus tips as the Kim at al. study predicts. Together, our results suggest a model that differs from the one proposed by Kim et al. (Kim et al 2007a). We believe that villus formation marks a transition from a multilayered epithelium that is not dependent on canonical Wnt signaling for proliferation to a single-layered pseudostratified epithelium that does require Wnt signals. We believe that the β-gal signal seen by the Shivdasani group between E16 and P3 is due to background staining but that the intervillus staining observed by this group at P3 and thereafter is probably real and is a part of the process of crypt formation. It would not be difficult to use in situ hybridization over these time points with probes for some additional Wnt targets (e.g., Cdx1, c-myc, CD-44v6) to further confirm our hypothesis. Also, a conditional (inducible) transgenic model that can inhibit Wnt signaling in the epithelium at late stages (E16 to P3) could be used to show that intervillus epithelial proliferation at this time is indeed reliant on canonical Wnt signals.

Although Wnt signaling was not modulated across the pyloric border, hedgehog signaling clearly was. There was a dramatic (7 fold) down-regulation of Shh in the intestine, but not the stomach at E16, and this was accompanied by a significant downregulation of the hedgehog target gene, Gli1. This difference was not present at E14. This is interesting because loss of Shh in the antrum causes that organ to look like intestine (Kim et al 2005b, Ramalho-Santos et al 2000). Also, constitutive

activation of Hh signaling in the intestine, at least in the frog model, causes failure of differentiation of the epithelial cells (Zhang et al 2001a). These observations suggest that levels of Hh might control the identity of the epithelium: high Hh retards intestinalization while lower Hh levels promote this process. Hh gradients also control cell type decisions in the neural tube, so this is consistent with the known functions of Hh signaling (Briscoe & Ericson 1999). The difference in the case of the intestine is that while the neural tube cells are direct targets of Hh signals (they express the receptor Patched and the transcription factor Gli1), the intestinal epithelial cells are not. Hh signaling in the intestine is strictly paracrine from epithelium to mesenchyme (Madison et al 2005). Therefore, the ability of Hh signals to control intestinal identity has to be due to crosstalk from the mesenchymal target cells. We could test this further by overexpressing Shh in the intestine using the villin promoter. We might be able to find some of the Shh target genes by examining gene regulation or performing ChIP on isolated mesenchyme after Shh administration.

It is important to realize that though the Hh pathway is turned down in the intestine, it is not turned off. Actually, other data from our lab shows that this lower level of expression is important for mesenchymal patterning of smooth muscle and innate immune cells in the adult (Lees et al., 2008; Kolerud et al., submitted).

We identified several border specific genes by comparing gene expression profiles of border tissues at E14.5 and E16.5 with those of stomach and intestine. Some of these are Gata3, Nkx2.5, nephrocan, and gremlin. Gata3 is a transcription factor and expressed in border mesenchyme; the finding of this gene at the border is novel, but we were helped by serendipity in the verification. The Engel lab, on the same floor of our building, was studying Gata3 expression and found an enhancer that drives expression of Gata3 right at the pyloric border. They also made reagents, like a Gata3

knockout/LacZ knockin and a Gata3-Cre that will help with further functional studies. They will soon publish a paper that shows that deletion of Gata3 at the pylorus results in its malformation. We would also like to see whether this affects intestinal maturation, but these studies have not yet been done.

Gremlin, a Bmp inhibitor, is expressed at E14.5 and E16.5 in border mesenchyme and the inner muscular layer of the duodenum. Previous studies showed that gremlin and Nkx2.5 are important for border formation (Moniot et al 2004, Theodosiou & Tabin 2003). Nephrocan is an inhibitor of the Tgfβ pathway and is expressed in border epithelium at E14.5 and at the base of forming antral glands and proximal intervillus duodenal epithelium at E16.5. This is a novel finding of our study. It is interesting that the two inhibitors (gremlin and nephrocan) are both expressed at the border. It will be important to find out what Bmp and Tgfβ receptors are nearby and to study Smad2/3 expression (a readout of Tgfβ signaling) and Smad 1/5/8 expression (a readout of Bmp signaling).

A situation where signaling proteins are expressed at the boundary between two tissues is also seen at the midbrain-hindbrain boundary. In that case, Fgf8 is expressed at the boundary and seems to be important for patterning tissue on both sides of the boundary (Canning et al 2007). This patterning action is called organizer activity. The cells of an organizer can even influence other cells when it is transplanted to another place. Though it is tempting to think that the pyloric boundary might act as an organizer to pattern the surrounding stomach and intestine, we have no evidence for this at this time. However, this is testable since we now have ability to culture isolated intestine and stomach for up to a week.

Our microarray experiments have taught us many new things about intestinal development, but this approach has some drawbacks too. Interpretation of microarray studies must be done with caution. Transcript levels do not necessarily accurately indicate protein expression or activity. In addition, that transcripts are not detected in a microarray study does not mean that gene is not expressed (No evidence is not the evidence of absence). Especially in microarray experiments using whole organs or multiple tissue types, as whole small intestine and stomach tissues in this thesis, the samples contain many cell types. Therefore, some genes may only be expressed in a few cells and their expression can be diluted to background level and are impossible to detect by microarray technology. For example, as mentioned above, Axin2, an important Wnt target gene that is often used to indicate Wnt pathway activity, is expressed at a very low level, close to background noise in our microarray experiment. When we did the in situ experiment, we found a very specific and interesting expression pattern for Axin2 that revealed important information about Wnt signaling in the maturing intestine. This example demonstrates that although the statistically best-supported list of up-regulated and down-regulated genes is important, the biological significance and wet lab experimental verification hold the final information about the genes.

An important and useful bioinformatics tool that we tried to take advantage of is promoter analysis and the identification of transcription factor binding sites (TFBS). Promoter analysis is an essential step on the way to identifying gene regulatory networks. Many algorithms have been developed to find TFBS matches and usually the number of sites detected is huge, in part because they often are short and not very stringent (see Chapter I introduction). Not all of the sites found are necessarily functional in the particular biological context. It would be very time consuming (but

accurate) to verify all the identified binding sites using traditional wet lab experiments, for example, gel shifting. In reality, however, using *in silico* strategies to examine such features as phylogenetic conservation (Doan et al 2004, Vanpoucke et al 2004), can dramatically reduce the number of candidate sites for testing.

Several studies have shown that gene expression is regulated by multiple transcription factors (Boehlk et al 2000, Fessele et al 2002, Werner et al 2003). Therefore, identifying higher order combinations of sites in promoters of co-regulated genes can also lead to a smaller set of more likely candidates for functional study. Genomatix has a platform for such an analysis, called the TFBS Framework analysis (Cartharius et al 2005, Cohen et al 2006, Scherf et al 2005, Seifert et al 2005). The identification of such motifs or models may imply synergistic or antagonistic effects for activating or repressing specific gene expression. The prerequisite for TFBS Frameworker analysis is knowing the candidate TFs and a small group of well-identified candidate target genes. Since there are over a thousand TFs in the mouse genome, the number of combinations is so huge that searching each possible combination is not feasible even with powerful computers. Therefore, one needs to identify potential transcription factors before searching the TFBS motifs. This could be accomplished by pathway analysis or literature mining, which can be done using BiblioSphere in Genomatix. The TFBS motif analysis performs best on small number of genes since large number of promoters will cause the combination problems similar with the situation of large number of TFs. A small group of potentially related (co-regulated) genes of can be obtained by clustering, GO term analysis, pathway analysis or in-depth literature mining. Once having the candidate promoters from this small group of genes, one can search the possible binding motifs characteristic of the identified TFs. The goal is to find patterns of binding motifs (for example, a site for TF#1 separated by 10 bp from

a site for TF#2) that are the same in multiple candidate promoter. After identification of such patterns, one can then search these patterns either in the whole promoter database or in the gene list identified from same microarray experiment to generate a second bigger gene list. If the function of this second list of genes is similar or highly related to the first small group of genes, the result makes sense from the biological standpoint. Otherwise, one needs to refine the original list of candidate TFs and/or genes and do the motif search again.

Although there are some successes with this strategy (Cohen et al 2006, Seifert et al 2005), we have extensively tried this method on groups of genes related to lipid metabolism or junctional complexes that we identified in the E16 duodenal epithelium. In the case of the lipid metabolism search, for example, excellent candidate transcription factors could be identified by literature mining using BiblioSphere, including Pparγ, Nr1h3, etc. and we knew that these TFs were also up-regulated in the D16 epithelial genes. Despite this, we could find no transcriptional patterns that appeared to be significant. There are several possible reasons for this. First, there may have been some false positive calls in the data. Second, our list is related to lipid metabolism which is a big term having many subgroups of terms and the genes related to this term may not all be regulated by the same group of transcription factors that we identified. If only a few of the large list of genes share a specific transcriptional framework, the inclusion of so many genes, requiring so many combinatorial searches, will limit the success of the search and tax the limited computer power of the Genomatix server. Third, the Genomatix strategies are centered on promoter regulation and the possible contributions of more distant sequence elements such as enchancers are ignored. To solve this problem, we could refine our gene list by pathway analysis to select only the most functionally

close-associated genes in one pathway map to reduce the number of genes searched. We did in fact try to do this, but still were not successful.

A large problem in this analysis is the fact that there are multiple transcriptional start sites (TSS) for each gene. For example, the human genome currently is annotated with 23,245 gene loci (NCBI 34) but there are 43,975 transcripts for these loci. About 45 percent (10,368) of the genes have alternative transcripts numbering from 2 to 40. Furthermore, 6,418 of the annotated loci have two or more promoters. Alternative transcripts of a gene are due several possible reasons: alternative splicing, alternative termination, or alternative first exons. The multiple TSS problem is a reflection of the various biological contexts in which a gene might be functionally involved. There is no way to identify which transcript is present or which TSS is used in which tissue using microarray expression data alone. Often, to avoid missing the "correct" promoter, we include them all in the analysis. But even when the gene list is small, this rapidly increases the search dimension. In a "needle in the haystack" problem, the last thing we need is to have a bigger haystack. It would be more ideal if this type of analysis could be done after some characterization of the 5' ends of transcripts that are actually expressed in the tissue. For example, we could combine the gene expression microarray data with exon array data or genome tiling array data to identify the functional transcripts for the tissue in question and then carry out the TFBS motif analysis. This would insure that only the relevant promoters are being searched, and will reduce the number of spurious TF patterns that are detected just due to the total length of the sequences searched.

Almost all of the analysis following the microarray experiments in this thesis is based on the identified gene list. As we discussed in the Chapter I, there are many ways to obtain these gene lists using different statistical packages. The different statistical

methods can of course result in different gene lists. Our early experience with the Affymetrix MAS4.0 package gave a surprisingly low concordance when the same data were analyzed using RMA. Although the RMA method that we used generates relatively reliable results, at least according to our PCR verification studies, we know that we may still miss some important genes whose expression level is very low or has high variation. A way to circumvent this problem is to use the entire microarray gene expression profiles and evaluate the gene expression systematically by focusing on the gene sets or pathways. One of the methods that we are exploring with our microarray data currently is Gene Set Enrichment Analysis (GSEA). This method uses only the raw gene expression data (not fold change and p-value) for determining enrichment of genes from a known pathway or function. We are comparing the same tissue at different times (dynamic comparisons, e.g. E14.5 stomach vs E16.5stomach) as well as different tissues at the same time (static comparisons, e.g. stomach vs. border at E14.5).

Another important aspect in microarray data analysis is meta-analysis by integration of multiple microarray datasets from multiple research groups (comparing the same tissues or different but closely related tissues or patient samples). There are many microarray databases available now and most publications require depositing the raw microarray data in some public database. One of the popular microarray databases is NCBI GEO (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/) (Barrett & Edgar 2006a, Edgar et al 2002). There are excellent publications about meta-analysis of GEO microarray data (Barrett & Edgar 2006b, Barrett et al 2007, Sean & Meltzer 2007). It would be interesting to integrate our microarray data with other gut-associated microarray data to verify or extend our own findings. We have done this in a small way by comparing our border array with arrays from

differentiating CaCo2 cells. These cells are thought to be a model cell line for the differentiation of intestinal cells. When they are plated at low density, they resemble undifferentiated intestinal epithelial cells. When they reach confluence, they up-regulate many genes associated with mature intestine and they polarize and become a tight monolayer. When we compared array data for CaCo2 differentiation (Saaf et al 2007) with our border array, we found excellent concordance of gene expression for many D16 epithelial genes. We believe that this means that we can use the CaCo2 cells as a model system to further understand the regulation of those concordant genes. Some genes, however, were dramatically up-regulated in our border array, but not in the CaCo2 array. We believe that these genes might require signals from the mesenchyme (which are missing in the CaCo2 cultures) to follow appropriate induction strategies.

A last point to make is that it is important to connect microarray data (transcriptomics) to proteomics. The proteins are the major functional molecule to fulfill most biochemical functions in organisms. Currently there are several databases and web servers for protein information including protein interactions. If we combine the protein-protein interaction data with gene expression data, we may build novel gene networks. This will give us even more power to build and test meaningful hypotheses about gut development. In summary, we want to know as much as possible about the significant genes or microarray data: biological function by integrating the data with Gene Ontology data base; characterized signaling transduction or metabolic pathways by combining with pathway databases (Genomatix, KEGG, BioCarta, TRANSPATH, etc); transcription regulation by Genomatix or TRANSFAC database, gene network by including known interaction between genes. These integrated analyses may help us in understanding what happened at the molecular level in the microarray experiment.

But it should be noted that it is rare that these analyses will result in one unifying hypothesis that can account for all the observed changes in microarray gene expression data. This is partly because most of our current knowledge of biomedical science comes from reductionistic approachs that study one gene, one protein or one pathway at a time and assume everything else was held constant during the experiment instead of currently massively parallel high-throughput methods in which a change in expression of one gene results in the changes of many other genes. We rarely have sufficient knowledge of a system to understand or explain why all these changes are occurring; this illustrates the importance of integrating of other available information. In addition, the statistical analyses that we currently use to identify differentially regulated genes assume that all genes are independent, which although helpful for the purposes of identifying candidate genes is an unrealistic simplification for complex biological systems. However, despite the knowledge that all genes are not independent, we currently lack the tools and methods for understanding which genes depend on each other. Once we have more information about gene interdependency, we will perform statistical analysis not on individual genes, but rather on networks of associated genes.

Bibliography

Barrett T, Edgar R. 2006a. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol* 411: 352-69

Barrett T, Edgar R. 2006b. Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*. *Methods Mol Biol* 338: 175-90

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, et al. 2007. NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res* 35: D760-5

Boehlk S, Fessele S, Mojaat A, Miyamoto NG, Werner T, et al. 2000. ATF and Jun transcription factors, acting through an Ets/CRE promoter module, mediate lipopolysaccharide inducibility of the chemokine RANTES in monocytic Mono Mac 6 cells. *Eur J Immunol* 30: 1102-12

Braunstein EM, Qiao XT, Madison B, Pinson K, Dunbar L, Gumucio DL. 2002. Villin: A marker for development of the epithelial pyloric border. *Dev Dyn* 224: 90-102

Briscoe J, Ericson J. 1999. The specification of neuronal identity by graded Sonic Hedgehog signalling. *Semin Cell Dev Biol* 10: 353-62

Canning CA, Lee L, Irving C, Mason I, Jones CM. 2007. Sustained interactive Wnt and FGF signaling is required to maintain isthmic identity. *Dev Biol* 305: 276-86

Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, et al. 2005. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics* 21: 2933-42

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 2: 65-73

Cohen CD, Klingenhoff A, Boucherot A, Nitsche A, Henger A, et al. 2006. Comparative promoter analysis allows de novo identification of specialized cell junction-associated proteins. *Proc Natl Acad Sci U S A* 103: 5682-7

Doan LL, Porter SD, Duan Z, Flubacher MM, Montoya D, et al. 2004. Targeted transcriptional repression of Gfi1 by GFI1 and GFI1B in lymphoid cells. *Nucleic Acids Res* 32: 2508-19

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30: 207-10

Fessele S, Maier H, Zischek C, Nelson PJ, Werner T. 2002. Regulatory context is a crucial part of gene function. *Trends Genet* 18: 60-3

Fleet JC, Wang L, Vitek O, Craig BA, Edenberg HJ. 2003. Gene expression profiling of Caco-2 BBe cells suggests a role for specific signaling pathways during intestinal differentiation. *Physiol Genomics* 13: 57-68

Garrison WD, Battle MA, Yang C, Kaestner KH, Sladek FM, Duncan SA. 2006. Hepatocyte nuclear factor 4alpha is essential for embryonic development of the mouse colon. *Gastroenterology* 130: 1207-20

Jho EH, Zhang T, Domon C, Joo CK, Freund JN, Costantini F. 2002. Wnt/beta-catenin/Tcf signaling induces the transcription of Axin2, a negative regulator of the signaling pathway. *Mol Cell Biol* 22: 1172-83

Kim BM, Buchner G, Miletich I, Sharpe PT, Shivdasani RA. 2005a. The stomach mesenchymal transcription factor Barx1 specifies gastric epithelial identity through inhibition of transient Wnt signaling. *Dev Cell* 8: 611-22

Kim BM, Mao J, Taketo MM, Shivdasani RA. 2007. Phases of canonical Wnt signaling during the development of mouse intestinal epithelium. *Gastroenterology* 133: 529-38

Kim JH, Huang Z, Mo R. 2005b. Gli3 null mice display glandular overgrowth of the developing stomach. *Dev Dyn* 234: 984-91

Luebke-Wheeler J, Zhang K, Battle M, Si-Tayeb K, Garrison W, et al. 2008. Hepatocyte nuclear factor 4alpha is implicated in endoplasmic reticulum stress-induced acute phase response by regulating expression of cyclic adenosine monophosphate responsive element binding protein H. *Hepatology* 48: 1242-50

Lustig B, Jerchow B, Sachs M, Weiler S, Pietsch T, et al. 2002. Negative feedback loop of Wnt signaling through upregulation of conductin/axin2 in colorectal and liver tumors. *Mol Cell Biol* 22: 1184-93

Madison BB, Braunstein K, Kuizon E, Portman K, Qiao XT, Gumucio DL. 2005. Epithelial hedgehog signals pattern the intestinal crypt-villus axis. *Development* 132: 279-89

Madison BB, Dunbar L, Qiao XT, Braunstein K, Braunstein E, Gumucio DL. 2002. Cis elements of the villin gene control expression in restricted domains of the vertical (crypt) and horizontal (duodenum, cecum) axes of the intestine. *J Biol Chem* 277: 33275-83

Moniot B, Biau S, Faure S, Nielsen CM, Berta P, et al. 2004. SOX9 specifies the pyloric sphincter epithelium through mesenchymal-epithelial signals. *Development* 131: 3795-804

Nechushtan H, Razin E. 2002. The function of MITF and associated proteins in mast cells. *Mol Immunol* 38: 1177-80

Partington GA, Fuller K, Chambers TJ, Pondel M. 2004. Mitf-PU.1 interactions with the tartrate-resistant acid phosphatase gene promoter during osteoclast differentiation. *Bone* 34: 237-45

Parviz F, Matullo C, Garrison WD, Savatski L, Adamson JW, et al. 2003. Hepatocyte nuclear factor 4alpha controls the development of a hepatic epithelium and liver morphogenesis. *Nat Genet* 34: 292-6

Ramalho-Santos M, Melton DA, McMahon AP. 2000. Hedgehog signals regulate multiple aspects of gastrointestinal development. *Development* 127: 2763-72

Saaf AM, Halbleib JM, Chen X, Yuen ST, Leung SY, et al. 2007. Parallels between global transcriptional programs of polarizing Caco-2 intestinal epithelial cells in vitro and gene expression programs in normal colon and colon cancer. *Mol Biol Cell* 18: 4245-60

Scherf M, Epple A, Werner T. 2005. The next generation of literature analysis: integration of genomic analysis into text mining. *Brief Bioinform* 6: 287-97

Sean D, Meltzer PS. 2007. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23: 1846-7

Seifert M, Scherf M, Epple A, Werner T. 2005. Multievidence microarray mining. *Trends Genet* 21: 553-8

Steingrimsson E, Tessarollo L, Pathak B, Hou L, Arnheiter H, et al. 2002. Mitf and Tfe3, two members of the Mitf-Tfe family of bHLH-Zip transcription factors, have important but functionally redundant roles in osteoclast development. *Proc Natl Acad Sci U S A* 99: 4477-82

Tachibana M. 2000. MITF: a stream flowing for pigment cells. *Pigment Cell Res* 13: 230-40

Theodosiou NA, Tabin CJ. 2003. Wnt signaling during development of the gastrointestinal tract. *Dev Biol* 259: 258-71

van de Wetering M, Sancho E, Verweij C, de Lau W, Oving I, et al. 2002. The beta-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. *Cell* 111: 241-50

Vanpoucke G, Goossens S, De Craene B, Gilbert B, van Roy F, Berx G. 2004. GATA-4 and MEF2C transcription factors control the tissue-specific expression of the alphaT-catenin gene CTNNA3. *Nucleic Acids Res* 32: 4155-65

Wang BE, Wang XD, Ernst JA, Polakis P, Gao WQ. 2008. Regulation of epithelial branching morphogenesis and cancer cell growth of the prostate by Wnt signaling. *PLoS ONE* 3: e2186

Werner T, Fessele S, Maier H, Nelson PJ. 2003. Computer modeling of promoter organization as a tool to study transcriptional coregulation. *FASEB J* 17: 1228-37

Yan D, Wiesmann M, Rohan M, Chan V, Jefferson AB, et al. 2001. Elevated expression of axin2 and hnkd mRNA provides evidence that Wnt/beta -catenin

signaling is activated in human colon tumors. *Proc Natl Acad Sci U S A* 98: 14973-8

Zhang J, Rosenthal A, de Sauvage FJ, Shivdasani RA. 2001. Downregulation of Hedgehog signaling is required for organogenesis of the small intestine in Xenopus. *Dev Biol* 229: 188-202

Zhang K, Shen X, Wu J, Sakaki K, Saunders T, et al. 2006. Endoplasmic reticulum stress activates cleavage of CREBH to induce a systemic inflammatory response. *Cell* 124: 587-99