12-12-2016 12:00 AM

Comprehensive Molecular Characterization of Human NODAL

Scott D. Findlay
The University of Western Ontario

Supervisor
Dr. Lynne-Marie Postovit
The University of Western Ontario

Graduate Program in Anatomy and Cell Biology
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Abstract

Nodal and related ligands are highly conserved members of the TGF-beta superfamily with well-established and essential roles in the early embryonic development of vertebrates, and in cell fate decisions in human embryonic stem (hES) cells. Aberrant NODAL signaling also generally promotes pro-tumourigenic phenotypes and the progression of a wide array of human cancers. Despite being pursued as a potential therapeutic target, many aspects of *NODAL*'s molecular biology remain poorly understood. This thesis provides a comprehensive characterization of gene expression from the human *NODAL* locus at multiple levels. First, an intronic *NODAL* SNP known as rs2231947 was found to be functional in its modulation of a novel alternatively spliced exon. This exon contributed to a full-length processed *NODAL* variant transcript. The existence of this genetically regulated *NODAL* isoform suggests that NODAL biology is more complex than currently appreciated. At the protein level, the alternatively spliced NODAL variant differs in the C-terminal half of the NODAL mature peptide. The NODAL variant was preferentially secreted relative to constitutive NODAL, but displayed similar extracellular stability and processing. Differential N-glycosylation was partially responsible for this increased secretion, and for NODAL secretion in general. The NODAL variant protein is unlikely to adopt a constitutive NODAL-like structure, and did not induce expression of targets of canonical NODAL signaling in the zebrafish embryo. However, the NODAL variant did efficiently complex via inter-chain disulfide bonds, and induced pro-tumourigenic phenotypes to a limited extent relative to constitutive NODAL. In summary, this work demonstrates previously unknown complexity governing human *NODAL* gene expression and function. These molecular details will help broaden our understanding of *NODAL* function as well as aid in the continued development of potential targeted therapies to inhibit NODAL signaling in cancer.

Keywords

NODAL, human embryonic stem cells, pluripotency, cancer, genetic heterogeneity, single nucleotide polymorphism (SNP), alternative splicing, alternative polyadenylation, N-glycosylation, precision genome editing.

Co-Authorship Statement

Each data chapter constitutes either a published manuscript, or a manuscript in preparation for submission. Author contributions are as follows:

Chapter 2: Characterization of a functional non-coding *NODAL* single nucleotide polymorphism (SNP).

This chapter consists of an extended version of:

Findlay, S. D., & Postovit, L.-M. (2016). Common Genetic Variation in Chromosome 10 q22.1 Shows a Strong Sex Bias in Human Embryonic Stem Cell Lines and Directly Controls the Novel Alternative Splicing of Human NODAL which is Associated with XIST Expression in Female Cell Lines. *STEM CELLS*, 34(3), 791–796. doi:10.1002/stem.2258

As the first author, I conducted all experiments and data analysis, and prepared the manuscript and figures.

Chapter 3: Characterization of human *NODAL* locus RNA variants

This chapter constitutes a manuscript in preparation for submission with the following authors:

Findlay, S.D., & Postovit, L.-M.

As the first author, I conducted all experiments and data analysis, and prepared the manuscript and figures.

Chapter 4: Function and post-translational regulation of NODAL proteins

This chapter constitutes a manuscript in preparation for submission with the following authors:

Findlay, S.D., Lypka, K., Waskiewicz, A.J. Postovit, L.-M.

As the first author, I conducted all experiments and data analysis, and prepared the manuscript and figures. K. Lypka assisted with *in situ* hybridizations and images for Figure 4.26.

Chapter 5: Differential effects of NODAL isoforms on cancer phenotypes, and improving NODAL modelling using precision genome editing

This chapter is an extended version of the NODAL-related data published in:

Findlay, S. D.*, Vincent, K. M.*, Berman, J. R., & Postovit, L.-M. (2016). A Digital PCR-Based Method for Efficient and Highly Specific Screening of Genome Edited Cells. *PLoS ONE*, 11(4), e0153901–17. doi:10.1371/journal.pone.0153901

As a first author, I conducted experiments and data analysis, and prepared the manuscript and figures. Specifically, I conceived of the study, selected all genome editing targets,

performed all plasmid cloning and modifications, genome editing of NODAL, as well as all design, validation, and performance of all ddPCR screening assays. Contributions from other authors appearing in this chapter are generation of the TALEN-edited cells screened in Figure 5.10 (K.M. Vincent), and assistance with all ddPCR assay designs (J.R. Berman). Additional genome editing of the SFRP1 locus performed by K.M. Vincent is not included in this thesis.

All functional assays in Figure 5.1 and PCR arrays in Figure 5.2 were performed by O. Bilyk and will be used in a separate manuscript.

Dedication

To my grandfather, David Findlay, who fostered my love for science, but was taken from us before he could see it flourish.

Acknowledgments

Thank you to all current and past members of the Postovit Lab at the University of Western Ontario and The University of Alberta. Thank you for your experimental and theoretical support of this project and for making the lab an enjoyable place to work and live. A special thank you to Guihua Zhang and Jiahui Liu for supporting and encouraging all students in the lab as if we were your own children. Another special thank you to Michael Jewer and Krista Vincent for the endless hours of debate and friendship we shared both in and out of the lab.

A huge thank you to Lynne Postovit for your support of my project, and more importantly, me. Thank you for the opportunity to work in your lab and to discover, explore, learn, and perhaps most importantly, sometimes struggle and fail. I have learned many valuable lessons and skills while working on this project, and I have you to thank for this. Most of all, thank you for your reassurance when needed, and for always putting up with my seemingly limitless skepticism and contrarianism.

Thank you, Alison Allan, for reading this thesis and for your thoughtful and thorough suggestions.

Thank you to my family for supporting this work. Specifically, my parents for the personal sacrifices they have made to afford me the privilege of doing this work and pursuing a career in science. Without you, none of this would have been possible.

Lastly, thank you Kim for the tremendous amounts of patience, support, and love that you always have for me. You kept me going, and this would not have been possible without you.

“In nature’s infinite book of secrecy

A *little* I can read.”

-William Shakespeare, *Antony and Cleopatra*. (emphasis added).

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

AAVS1	Adeno-Associated Virus integration Site 1
Alu	Arthrobacter luteus
AMH	Anti-Mullerian Hormone
APA	Alternative Polyadenylation
AS	Alternative Splicing
ASE	Asymmetric Enhancer
AVE	Anterior Visceral Endoderm
BLAST	Basic Local Alignment Search Tool
BMP	Bone Morphogenic Protein
Cas	CRISPR-associated
CHD	Congenital Heart Defect
ChIP	Chromatin Immunoprecipitation
CM	Conditioned Media
CRISPR	Clustered Regularly Interspersed Short Palindromic Repeats
dbSNP	Single Nucleotide Polymorphism Database
ddPCR	droplet digital Polymerase Chain Reaction
DNA	Deoxyribonucleic Acid
DSE	Downstream Sequence Element
DVE	Distal Visceral Endoderm
EGF-CFC	Epidermal Growth Factor-cysteine-rich Cripto-1/FLR1/Cryptic
EMT	Epithelial-to-Mesenchymal Transition
ENCODE	Encyclopedia of DNA Elements

ER	Endoplasmic Reticulum
ERE	Epigenetic Regulatory Element
GDF	Growth Differentiation Factor
gDNA	genomic Deoxyribonucleic Acid
GFCK	Growth Factor Cystine Knot
GPI	Glycosylphosphatidylinositol
GWAS	Genome-Wide Association Study
HBE	Highly-Bound Element
hES	human Embryonic Stem Cell
HPE	Holoprosencephaly
hPSC	human Pluripotent Stem Cell
iPS	induced Pluripotent Stem Cell
ISCI	International Stem Cell Initiative
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MAPK	Mitogen-Activated Protein Kinase
MEF-CM	Mouse Embryonic Fibroblast-Conditioned Media
mEpiSC	mouse Epiblast Stem Cell
mES	mouse Embryonic Stem Cell
MET	Mesenchymal-to-Epithelial Transition
mRNA	messenger Ribonucleic Acid
NAT	Natural Antisense Transcript
NCBI	National Center for Biotechnology Information
NDE	Node-Specific Enhancer

NHEJ	Non-Homologous End Joining
NMD	Nonsense-Mediated Decay
ORF	Open Reading Frame
PAS	Polyadenylation Signal
PBS	Phosphate-Buffered Saline
PDB	Protein Data Bank
PEE	Proximal Epiblast Enhancer
PKC	Protein Kinase C
PTC	Premature Termination Codon
PTM	Post-Translational Modification
RACE	Rapid Amplification of Complementary Ends
RT	Reverse Transcription/ Reverse Transcriptase
SINE	Short Interspersed Elements
SNP	Single/Simple Nucleotide Polymorphism
snRNP	small nuclear Ribonucleoprotein
TALEN	Transcription Activator-Like Effector Nuclease
TGF-beta	Transforming Growth Factor beta
UCSC	University of California Santa Cruz
USE	Upstream Sequence Element
UTR	Untranslated Region
VEGFA	Vascular Endothelial Growth Factor A
Xnr	Xenopus nodal-related

Chapter 1

1 Introduction and literature review

1.1 The cancer problem

As the leading cause of death in Canada, cancers of various forms were responsible for an estimated 78,000 deaths in 2015 [1]. The term "cancer" describes a collection of cellular pathologies, originating in a multitude of organs as diverse as the skin, lungs, and brain. Regardless of their origin and etiology, cancers share certain hallmarks including uncontrolled cell growth, resistance to therapy, and the ability to spread systemically and to other organs in a process known as metastasis [2, 3]. Research in the last half century has shown that these hallmarks are manifestations of widespread genomic DNA alterations ranging from point mutations to gross chromosomal abnormalities including duplication or deletion of entire chromosomes. Since genomic DNA serves as the template for all gene expression, widespread DNA mutation and genomic instability drastically alters cellular behaviour. Thus, in a broad sense, cancers involve disruptions of normally exquisitely regulated cellular and sub-cellular processes.

1.2 Modelling cancer biology *in vitro*

Human cancers are generally studied through two complementary approaches: 1) Analyzing clinical specimens such as tumour biopsies, and 2) The establishment of model systems amenable to experimental manipulation. These include tumour cells adapted for culture *in vitro* as cell lines, and the propagation of human cancer cells as tumour xenografts in animal models such as mice.

1.3 The cancer stem cell hypothesis and phenotypic plasticity

As a population, tumour cells have an uncanny ability to withstand an onslaught of host defenses including cell cycle blockade and apoptosis normally invoked in response to DNA damage, and targeted cytotoxicity against transformed cells by host immunosurveillance mechanisms [2, 3]. Beyond surviving these initial transforming events, a highly proliferative tumour must also meet the rapidly increasing demands for

both energy production and biosynthesis of cellular materials. This involves a dramatic shift in cellular metabolism to favour glycolysis in what has been described as "reprogramming energy metabolism" (reviewed in [2, 3]). In malignant tumours, those cells that do metastasize beyond the site of the primary neoplasia must also survive relatively harsh environments including the bloodstream and lymphatic circulation, as well as other organs where they may lack support provided by other tumour cells as in a primary tumour. Indeed, multiple steps in the metastatic cascade, especially colonization, are very inefficient in experimental models of metastasis [4]. Beyond these challenges intrinsic to their natural environment, tumours can also withstand various toxic chemotherapies deployed as a major part of cancer patient treatment regimens. While such treatments are generally effective in eliminating a great majority of the tumour mass, a very small number of cancer cells almost invariably survive. These cells eventually contribute to patient relapse manifested by a thriving tumour that is often now highly resistant to the initial treatment [5].

There are two general and non-mutually exclusive aspects of tumour biology that account for this resilience. First, the cancer stem cell model suggests that there is a subpopulation of tumour cells with the ability to clonally regenerate an entire tumour. These cells are self-renewing and can also give rise to a heterogeneous tumour (reviewed in [6]). Whether the genesis of these stem-like cells is a stochastic process, or they are a biologically distinct cell type at the "top" of a tumour cell hierarchy has been the source of great debate (e.g. [7-10]). Regardless of their origin or exact nature, these cells are thought to be imperative for maintenance of tumour growth, seeding of metastases, and resistance to therapy. Therefore, it is unsurprising that cancer stem cells have received a great deal of attention and hold much promise as viable targets in the next generation of precision cancer therapy development.

Behaviourally, a cancer stem cell, and likely other tumour cells, must be able to respond to external cues in order to promote the appropriate cellular behaviour required for propagation. Of course, this requires signal transduction pathways and other sub-cellular machinery to be intact, despite a high mutation load and general genetic instability. As an example, a tumour cell may sense a lack of oxygen and respond by secreting pro-

angiogenic growth factors such as vascular endothelial growth factor A (VEGFA/VEGF) to promote recruitment of host endothelial cells and subsequent local blood vessel extension through angiogenesis [11]. This ability to respond to microenvironmental cues can consist of much more complex cellular responses including changes in cell identity, in a process that can generally be referred to as "phenotypic plasticity." Phenotypic plasticity is generally potentiated by a reversible change in epigenetic state(s) (reviewed in [12]). Perhaps one of the most well-studied examples of this plasticity is the ability to undergo an epithelial-to-mesenchymal transition, or "EMT." This process is a major driver of the carcinoma cell's ability to escape the primary tumour, invade through a basement membrane, and enter the circulation for potential seeding of secondary metastases. Although a hallmark process in cancer, EMT is actually a normally occurring process at numerous stages of early embryonic development (reviewed in [13], and was first characterized in the primitive streak of a chick blastocyst [14]—a structure that initiates germ-layer formation and sets the stage for gastrulation. While initially referred to as epithelial-to-mesenchymal *transformation*, it has more recently been dubbed a *transition*, after the EMT process was remarkably shown to be reversible in the form of a mesenchymal-to-epithelial transition (MET) [15]. Furthermore, there is now evidence of a cancer-specific partial EMT hybrid phenotype [13, 15, 16]. These transitions are hallmark examples of extreme phenotypic plasticity afforded to cancer cells through the "hijacking" of normal cellular processes out of their appropriate contexts. This exploitation is the very rationale for the study of non-cancerous models of normal stem cell biology to inform our understanding of human cancer.

1.4 Human embryonic stem cells

Non-human models such as zebrafish, xenopus, mouse, and chicken are commonly used to study early embryonic development in complete biological systems. Owing to the relatively high degree of genetic relatedness between humans and these model organisms, as well as evolutionary constraints on embryonic development, a great deal of our basic understanding of embryology gleaned from model organisms is generally applicable to human development. However, due to the practical and ethical limitations of studying early human embryonic development *in utero*, *direct* study of early human development

has generally been limited to established human pluripotent stem cell (hPSC) lines, as well as surplus embryos from *in vitro* fertilization processes. The first hPSCs to be derived were human embryonic stem (hES) cells [17]. These cells were isolated from the inner cell mass of pre-implantation blastocysts and adapted to *in vitro* cell culture. Human ES cells are both self-renewing and able to ultimately derive the full panoply of adult cell types—a property known as pluripotency. More recently, human induced pluripotent stem (iPS) cells have been derived from human adult somatic cells such as skin fibroblasts [18]. These cells are generally reprogrammed with the "Yamanaka factors" POU5F1 (OCT4), SOX2, KLF4, and MYC, to activate and reinforce core regulatory gene expression networks for pluripotency [18]. Successfully reprogrammed iPS cells are functionally equivalent to hES cells as they are both self-renewing and pluripotent. Induced pluripotent stem cells are extremely useful for modelling stem cell biology across different genetic backgrounds of interest. Accordingly, iPS cells have been praised for their potential applications in personalized and regenerative medicine, and have also been used as cancer models to enhance patient and cancer-specific disease modelling [19].

Unlike their previously derived mouse embryonic stem (mES) cell counterparts, it has been suggested that hES cells continue along their developmental trajectory during derivation from pre-implantation inner cell mass cells, and share many features with post-implantation embryos, including X inactivation in female cells, and high expression of genes related to NODAL/Activin signalling [20]. Furthermore, hES cells share several characteristics with mouse epiblast stem cells (mEpiSC) subsequently derived from post-implantation embryos [21, 22]. These include flat colony morphology, inefficient single-cell cloning ability, and reliance on similar signalling pathways [23, 24]. Further work has demonstrated that hES cells and EpiSCs exist in an epigenetically "primed" pluripotent state poised for differentiation, defined in part by bivalent histone marks. Both the inhibitory H3K27me3 histone modification and the activating H3K4me3 histone modification are found at promoters of genes to be transcribed as part of an early differentiation response in hES cells [25]. Notably, additional histone marks are also important in maintaining the pluripotent state (reviewed in [26]), and bivalent chromatin marks have also been characterized in mouse ES cells [27]. Human "naive" ground-state

pluripotent stem cells have subsequently been derived that share features with mouse ES cells [28, 29]. Thus in both mouse and human, there are multiple distinct pluripotent stem cell states that can be modelled *in vitro* [30].

1.5 The transforming growth factor (TGF)-beta superfamily

Research aimed at elucidating the mechanisms by which pluripotency is maintained and cell fate choice is made in early embryonic stem cells has uncovered signalling by TGF-beta superfamily members as a major regulator of the epigenetic changes governing these processes [31-34]. More generally, the TGF-beta gene family plays major and complex roles in early embryonic development, stem cell biology, and cancer. The TGF-beta superfamily contains at least 30 members in humans, and is well conserved across vertebrates, with its family members playing important roles in a myriad of cellular processes including cellular differentiation, proliferation, and migration in a wide variety of cell types and contexts (reviewed in [35]). Classes of proteins that constitute the superfamily include the TFG-betas themselves, bone morphogenic proteins (BMPs), activins, growth and differentiation factors (GDFs), and other members such as anti-Mullerian hormone (AMH) and nodal growth differentiation factor (NODAL).

Members of the superfamily generally share similar structures including an N-terminal signal peptide for secretion, an adjacent pro-domain, and a C-terminal peptide cleaved from the pro-domain to yield mature and active ligand. Family members also contain a cystine knot motif consisting of three or four intrachain disulfide bonds that follow the growth factor cystine knot (GFCK) pattern. The TGF-beta protein represents one of four prototypical GFCK structures [36]. The participating cysteines provide TGF-beta characteristic structure and are similarly positioned across family members. Additionally, TGF-beta proteins utilize an additional cysteine that does not participate in the cystine knot to form homodimers or heterodimers with other related proteins. Structurally, TGF-beta proteins consist of a helical “wrist” with two beta-sheet-rich “finger” domains extending outward. The finger domains form a pocket for the wrist domain of the dimerizing ligand, with the interchain disulfide bond at the centre of the structure (Figure 1.1).

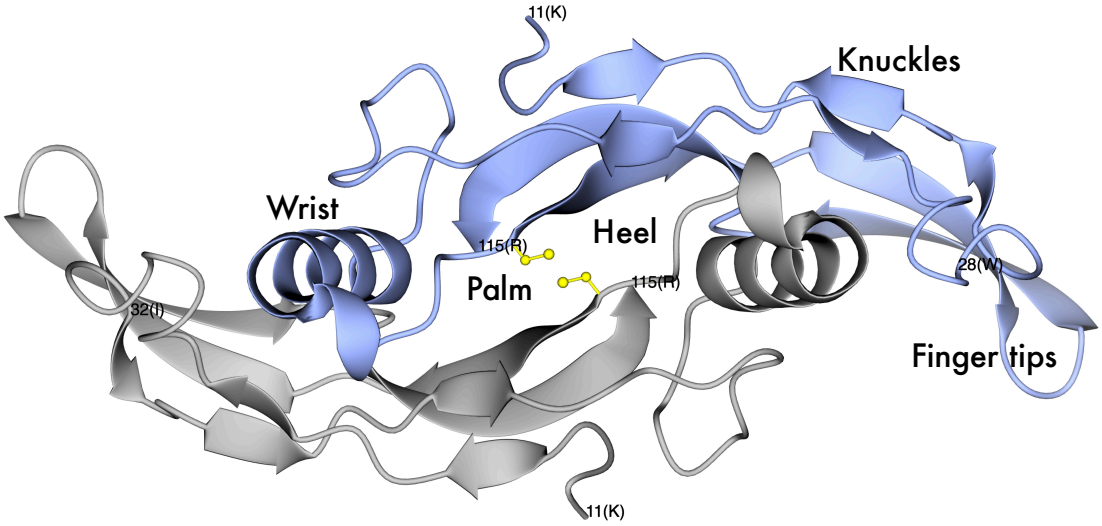


Figure 1.1: A NODAL chimera homodimer illustrating TGF-beta superfamily structure.

Each polypeptide subunit is coloured separately. Descriptions for portions of the structure are shown for the top subunit only. Side chains of the interchain disulfide bond-forming cysteine residues are shown in yellow.

In terms of signalling, extracellular TGF-beta ligands bind heterodimeric complexes consisting of one of seven type I and one of five type II serine/threonine kinase receptors. Upon ligand binding, receptor complex formation triggers phosphorylation of the type I receptor via kinase activity of the constitutively active type II receptor [37, 38]. Additional membrane-bound co-receptors also modulate ligand-induced signalling for some family members. Some co-receptors are essential for downstream signalling [39], while others are enhancing or even inhibiting [40]. Upon activation, the type I receptor directly phosphorylates intracellular proteins known as mediator Smads [41]. Fascinatingly, and seemingly contrary to the diversity of receptor-ligand interactions in the TGF-beta superfamily, downstream of receptor activation, signals from most family members converge on one of two main groups of mediator Smads [42]. These are the BMP-activated Smads (Smad1, Smad5, Smad9), and the TGF-beta-activated Smads (Smad2, Smad3) (reviewed in [35]). Phosphorylation of mediator Smads potentiates their association with the common mediator Smad, Smad4. Complex formation promotes their accumulation in the nucleus where Smad complexes regulate gene expression of various target genes in cooperation with DNA binding proteins and other co-repressors or co-activators [43-45]. TGF-beta superfamily members have also been shown to induce Smad-independent signalling events including activation of the MAP kinase pathway [46]. In addition to conventional serine/threonine kinase activity, superfamily receptors can also display limited tyrosine kinase activity[47]. Lastly, there are various points of direct cross-talk between TGF-beta signalling and other singling pathways such as the Wnt signalling cascade [48].

1.6 The TGF-beta superfamily member NODAL

One of the aforementioned TGF-beta superfamily members is “nodal growth and differentiation factor” in human (gene symbol: *NODAL*, NCBI gene ID: 4838), and the closely related “nodal” in mouse (gene symbol: *Nodal*, NCBI gene ID: 18119). *Nodal* is aptly named after its discovery in the mouse node [49], a cluster of cells at the distal end of the primitive streak in gastrula-stage embryos [50]. *Nodal* is a gene with numerous essential roles in early development, and it has been well studied in numerous vertebrate embryos and *in vitro* models of early development and stem cell biology [51-54].

1.7 Nodal signalling

Signal transduction initiated by extracellular Nodal ligands is illustrated in Figure 1.2. Nodal is secreted as a pro-protein where it is generally extracellularly cleaved by the proteolytic activities of secreted pro-protein convertases Furin and Pcsk6 (also known as Pace4) [55]. The resultant mature Nodal peptides can homo-dimerize to engage both type I tyrosine kinase receptors Alk4 or Alk7 (also known as Acvr1B and Acvr1C, respectively), and type II receptors Acvr2A or Acvr2B (formerly known as ActrIIa and ActRIIB, respectively) (reviewed in [35]). Two glycosylphosphatidylinositol (GPI)-linked and membrane bound members of the epidermal growth factor-cysteine-rich Cripto-1/FLR1/cryptic (EGF-CFC) family serve as requisite co-receptors for Nodal signals. Cripto or Cryptic bind the type I Nodal receptor and help recruit type II receptors to facilitate a functional receptor complex [56]. Interestingly, although Cripto is generally required for Nodal signalling, Cripto-independent signalling has been described in the mouse embryo [57, 58]. Complete receptor complex formation triggers Nodal signal transduction through phosphorylation of mediator Smads Smad2 and Smad3 and their subsequent interaction with Smad4 to facilitate nuclear translocation. In the nucleus, Smad complexes interact with transcription factors such as forehead box 1 (FoxH1) to drive transcription of target genes including *Gsc* [59], as well as *Nodal* itself. This positive feedback transcriptional response is facilitated by a Smad2/FoxH1-bound enhancer in intron 1 of *Nodal* [60]. Other transcriptional targets of Nodal signalling include Lefty. Lefty proteins are secreted endogenous inhibitors of Nodal signalling. Direct binding of Lefty to either Nodal or Cripto/Cryptic can prevent successful receptor ligand complex formation [61]. Simultaneous upregulation of both agonists and antagonists of Nodal signalling suggests that Nodal signals are carefully regulated during embryonic development. Indeed, the spatiotemporal regulation of Nodal signalling needs to be carefully balanced and precisely regulated as cells are sensitive to both the dose and duration of Nodal signals [62]. Co-expression of both Nodal and Lefty takes advantage of differential diffusion of these two secreted proteins, with the more stable Lefty protein restricting Nodal expression far from the source, whereas short range Nodal signals are more potent [63-65]. However, the pervasiveness of this effect was recently challenged by the finding that a short-range temporal “window” of Nodal-related expression was

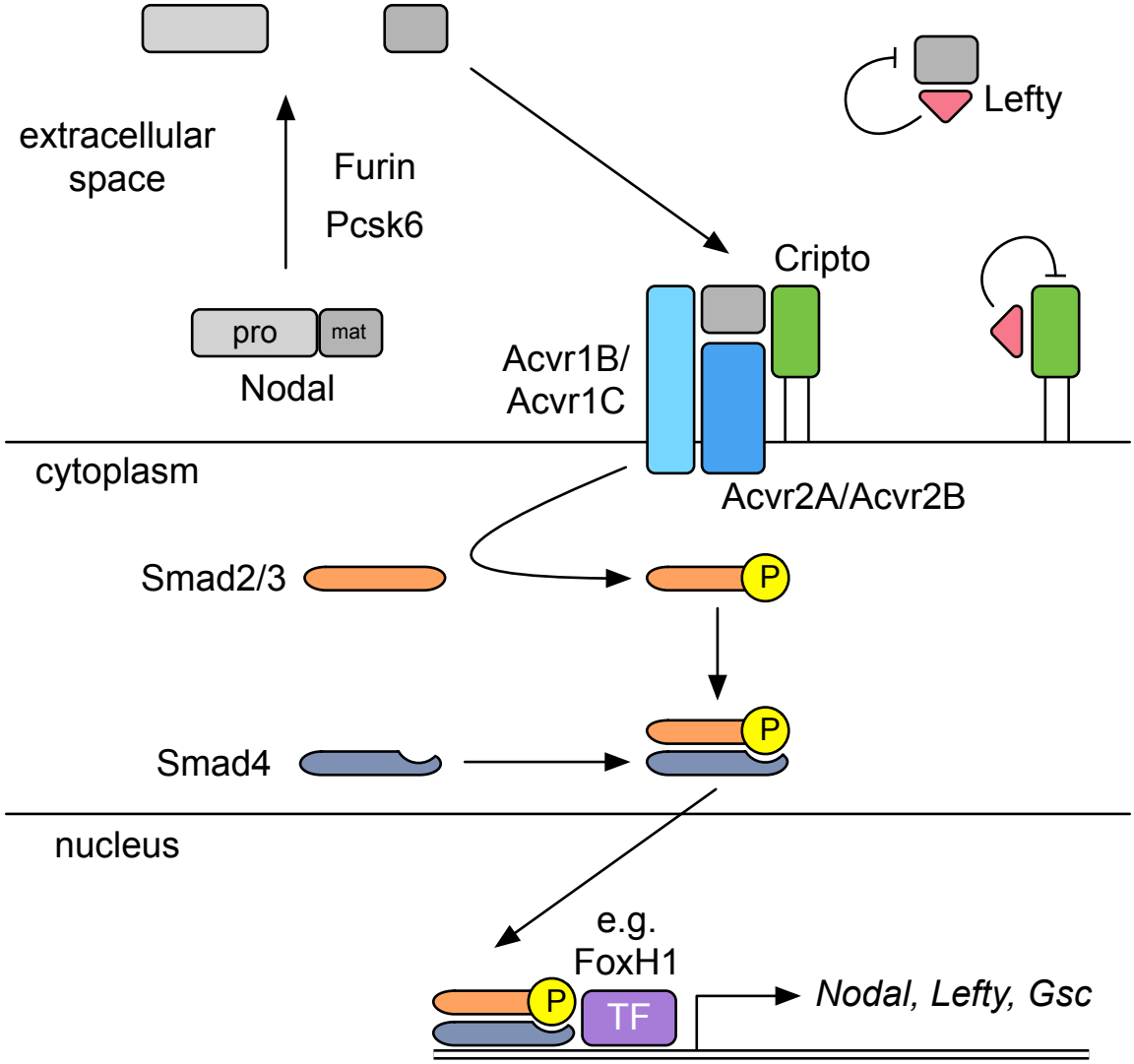


Figure 1.2: A schematic of Nodal signal transduction.
Nodal is extracellularly processed and facilitates receptor complexing, activating type I receptors Acvr1B or Acvr1C that directly phosphorylate intracellular Smad2/3 proteins. Phosphorylation facilitates their interaction with Smad4, forming complexes that are able to translocate to the nucleus. In the nucleus, these complexes interact with transcription factors to transcribe target genes such as Nodal itself and the Nodal inhibitor Lefty. Extracellular Lefty can inhibit Nodal signaling through interactions with Nodal or Cripto, preventing proper signalling receptor complex formation. “pro” indicates the N-terminal Nodal pro-domain/peptide. “mat” indicates the C-terminal mature Nodal domain/peptide. “TF” = transcription factors. Pointed arrows indicate activation. Blunt arrows indicate inhibition.

sufficient to establish a Nodal signalling gradient in the zebrafish embryo, and that the duration of this window was regulated by micro RNA-mediated translational repression of *Lefty* [66]. Other endogenous Nodal inhibitors outside of Lefty known as Cerberus proteins also bind Nodal directly to inhibit receptor-ligand interactions [67].

1.8 Nodal in the developing mouse embryo

Much of our general understanding of *Nodal*'s role in embryonic development comes from the study of mouse embryos. In blastocyst stage embryos, *Nodal* expression is high in the epiblast and promotes expansion of this structure while preventing spontaneous differentiation [68]. Following implantation, *Nodal* is expressed throughout the epiblast and contributes to specification of the distal visceral endoderm (DVE; [69]). In turn, the DVE secretes Nodal inhibitors, establishing a proximal-distal Nodal gradient. This morphogen gradient is one of the first in the developing embryo and helps define the first embryonic axis to develop [70]. Nodal subsequently directs the DVE toward the prospective anterior side of the embryo where it is now termed the anterior visceral endoderm (AVE), and contributes to definition of the anterior-posterior axis [70, 71]. Collectively, these coordinated events result in the restriction of *Nodal* expression to the proximal posterior epiblast prior to the onset of gastrulation. Through interactions with the extra-embryonic ectoderm, *Nodal* is amplified in the epiblast as a requisite to initiate primitive streak formation [72, 73]. The primitive streak is established on the posterior side of the embryo. As epiblast cells invaginate into the streak from the proximal end of the embryo, they undergo EMT. Those that migrate laterally are exposed to a relatively low Nodal dose and become mesoderm. Those that continue to migrate toward the distal end receive a high dose Nodal signal specifying definitive endoderm (reviewed in [54]). These cells form the primitive node, around the periphery of which *Nodal* is expressed [49]. This structure is an organizing centre for establishing left-right asymmetry of organ development in vertebrates [74]. To the left of the node, Nodal activity is relatively high in the left lateral plate mesoderm, while Nodal activity is restricted by expression of Lefty in the right lateral plate mesoderm [75-77].

Unsurprisingly, these *Nodal* functions are imperative to proper development. Specifically, *Nodal*$^{-/-}$ mutation is embryonic lethal in mice [49, 73], likely due to the

failure of these embryos to form the primitive streak for the initiation of gastrulation [72]. However, *Nodal* $^{wt/-}$ mice develop normally [78, 79], suggesting that either lower levels of Nodal are sufficient, or that the embryo naturally compensates for reduced *Nodal* expression. Furthermore, mice with hypomorphic *Nodal* allele(s) display partial lethality and a spectrum of developmental defects concerning heart and brain development, as well as laterality [80, 81].

1.9 NODAL in human pluripotent stem cells

Owing to practical and ethical limitations concerning research on human embryos, human-specific study of *NODAL* biology has generally been limited to cultured hES cells. In hES cells, *NODAL* helps maintain pluripotency [82], and block differentiation toward neuroectoderm lineages [83], in part through positive regulation of NANOG expression [84]. Transcriptional changes driving cell fate decisions are mediated by nuclear complexes containing active SMAD2/3 [32-34, 84]. NODAL/Activin signalling is also involved in the deposition of activating H3K4me3 marks at gene promoters [31]. A role for *NODAL* in both the maintenance of stem cell pluripotency and the promotion of mesendoderm differentiation as described above may seem paradoxical. However, the dose and duration of NODAL signal (reviewed in [62]), as well as the presence of SMAD2/3 complexing proteins such as NANOG [34] are important in determining how NODAL affects cell fate. Thus, it is apparent that *NODAL* plays several distinct roles in early embryonic development, and context is very important in dictating *NODAL* function. The next section will detail the impact of *NODAL* expression in human cancers, where normal developmental contexts are all but lost.

1.10 The impact of NODAL expression in human cancers

NODAL expression in cancer was first identified by Postovit and Topczewska and colleagues in the aggressive C8161 human melanoma cell line [85]. These cells were able to induce ectopic outgrowths or a complete secondary axis after injection into zebrafish embryos at the blastocyst stage. *NODAL* was identified as the primary factor responsible for this induction as inhibition of NODAL signalling by LEFTY1, and reduction of NODAL levels using a *NODAL* antisense oligonucleotide morpholino both abrogated

C8161-induced outgrowth. Clinically, immunohistochemistry revealed NODAL protein was present in human metastatic melanoma tissues, but not in normal skin. Experimentally, inhibition of *NODAL* in C8161 cells reduced anchorage-independent growth capacity in a soft agar colony formation assay, as well as tumour growth in a mouse xenograft model. Since this pioneering discovery, *NODAL* expression has been shown to affect numerous tumour phenotypes in experimental models of several human cancers including cancers of the breast [86-90], prostate [91, 92], ovary [93, 94], and pancreas [95], as well as glioma [96, 97], glioblastoma [98], endometrial cancer [99], hepatocellular carcinoma [100], and choriocarcinoma [89, 101, 102]. In these models, *NODAL* impacts numerous processes including phenotypic plasticity, proliferation and apoptosis, migration and invasion, EMT, angiogenesis, and metastasis (reviewed in [53]). In general, a pro-tumourigenic role for NODAL has been shown, although a notable minority of studies have demonstrated decreased proliferation and increased apoptosis resulting from NODAL signaling [93, 103].

There is also strong correlative evidence of a link between high *NODAL* expression and poor clinical outcome in numerous cancers. Prominent examples include a study of over 400 breast cancer patients where NODAL correlated positively with tumour stage and grade, independently of estrogen receptor/progesterone receptor (ER/PR) or HER2 status [104]. Moreover, a recent meta-analysis of *NODAL* expression in human cancers originating from 11 different tissues and including more than 800 cancer patients revealed significantly higher expression in cancerous tissue relative to healthy control tissue. A subset of studies analyzed also revealed significant positive correlations between *NODAL* expression and high tumour grade (III & IV relative to I & II) and tumour size, and a significant negative correlation between *NODAL* expression and degree of differentiation [105].

After fulfilling its early embryonic functions, *NODAL* is epigenetically silenced at least in part through polycomb repressive complex-mediated H3K27me3 deposition at the *NODAL* locus [106]. In adults, *NODAL* expression is thought to be generally limited to select niches including the mammary gland, cycling endometrium, adult liver stem cells, and pancreatic beta cells (reviewed in [53]). Still, in most adult tissues, *NODAL*

expression is silenced in normal tissue. During cancer progression, *NODAL* expression is activated via mechanisms that are not yet well understood [53].

Notably, despite coordinated induction of *NODAL* and Lefty gene expression in development, low or undetectable levels of Lefty have been reported in several cancers in both patient samples and cell lines [99, 107], suggesting that NODAL is not always subject to this mechanism of endogenous inhibition in cancer. Another study has also shown that NODAL can engage non-canonical receptor complexes in cancer but not in hES cells [108]. These are examples of how the context dictating NODAL function can differ dramatically between evolutionarily constrained and carefully-regulated developmental systems, and much more chaotic and deregulated cancerous systems. These and other not yet discovered differences are important to consider when studying NODAL function in cancer, and are further complicated by inter-tumour and intra-tumour heterogeneity.

1.11 Inhibition of NODAL activity as a targeted cancer therapeutic strategy

Inhibitors of components of the *NODAL* signalling pathway are currently being developed for targeted cancer therapy. These consist of a monoclonal antibody targeting Cripto-1 [109], and an inhibitor with activity against Alk4/7 [110]. More recently, encouraging pre-clinical results have been reported for a newly developed monoclonal antibody termed 3D1 that targets NODAL protein directly [111, 112]. This antibody was developed against the pre-helical loop region of mature NODAL implicated in Cripto-1 binding. Treatment of NODAL-expressing cancer cells with 3D1 recapitulated many of the previous effects of NODAL inhibition, including reduced phosphorylation of SMAD2 and ERK. In mouse a mouse xenograft model, treatment with 3D1 resulted in reduced tumour growth and reduced metastatic potential. Notably, most of these results were demonstrated for the C8161 melanoma cell line. It will be of interest to see if future preclinical modelling of NODAL inhibition by 3D1 is robust across different cancer types. Additionally, it was not demonstrated if the effects of 3D1 were specific to NODAL. Regardless of the strategy used to inhibit NODAL signalling, successful development of any targeted therapy depends on a detailed understanding of the target

molecule.

1.12 Direct study of human NODAL is lacking

Like many genes, the molecular complexity of *NODAL*'s regulation and expression are often either overlooked or difficult to incorporate into conventional experimental model systems. Furthermore, a great deal of molecular and functional knowledge of *NODAL* has been obtained from study of non-human embryos and stem cells. Thus, while model systems such as the mouse embryo have been extremely valuable for studying *NODAL*, this knowledge must be supplemented with data from human models to fully appreciate its human-specific role in development and cancer pathology. Indeed, considerable differences in development exist between species as divergent as mouse and human. *NODAL* is no exception, as differences in *NODAL* biology between human and mouse ES cells have been described [113]. In addition, many aspects of *NODAL* biology are also inferred from similar superfamily members such as TGF-beta, Activin, and the GDFs. For example, Alk4/5/7 receptors transmit signals from several TGF-beta superfamily members, but inhibition of these receptors with the small molecule inhibitor SB-431542 [114] is often used to infer NODAL function, although the inhibitor is not specific to this ligand (e.g. [82, 84, 115]). As another example, the NODAL cysteines ostensibly involved in disulfide bond formation and homo-dimerization are annotated by similarity to other superfamily members, and have not been directly studied.

1.13 Transcriptional regulation of gene expression

For a typical protein coding gene, regulation of its expression takes place at numerous stages. Transcription is perhaps the most well-studied point of regulation for many protein coding genes. Transcription is generally governed by recruitment of transcription factors to enhancer elements that associate with the basal RNA polymerase II machinery to initiate transcription. Epigenetic contexts such as DNA methylation and post-translational histone modifications help to modulate transcription at a given locus [116]. Control over *Nodal* transcription has been particularly well studied, again mainly in mouse systems [117, 118]. Characterized enhancer elements influencing *Nodal* transcription are shown in Figure 1.3. In the mouse embryo, Nodal controls its own

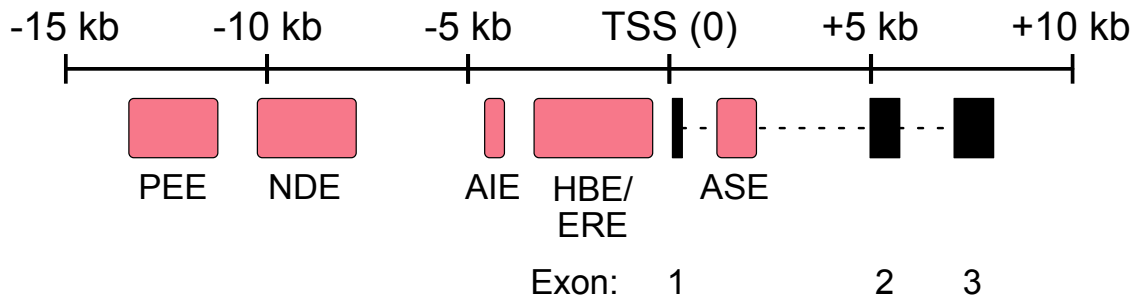


Figure 1.3: Mouse *Nodal* enhancers.
The approximate genomic locations of characterized *Nodal* enhancers (pink) are shown relative to *Nodal* exons (black; untranslated regions included). Numbers indicate approximate coordinates relative to the *Nodal* transcriptional start site ("TSS"). The dashed lines indicate introns. "PEE" = proximal epiblast enhancer. "NDE" = node-specific enhancer. "AIE" = asymmetric initiator element. "HBE/ERE" = highly bound element or epigenetic regulatory element. "ASE" = asymmetric enhancer. Locations of all elements are approximate.

expression through a positive feedback loop mediated by a FoxH1-responsive asymmetric enhancer (ASE) element in the first intron of *Nodal* [60, 118]. Other characterized *Nodal* enhancer elements include a node-specific *Nodal* enhancer (NDE; [74, 118-120], a proximal epiblast enhancer (PEE;[118, 121]), and an asymmetric initiator element/left side specific enhancer (AIE;[122]).

More recently, an enhancer termed the "highly bound element" (HBE) has been described [123]. This element is required for *Nodal* expression in the mouse epiblast and drives *Nodal* expression in an Oct4-dependent manner. Furthermore, previous studies had identified the HBE locus as a multi-transcription factor-binding locus in ES cells [124-126]. Transcription factors found to be bound to this element include the master regulators of pluripotency Oct4, Nanog, and Sox2. Another study also termed this region the "epigenetic regulatory element" (ERE) and found it was subject to DNA methylation that regulated *Nodal* expression and Oct4 binding [127]. Although our understanding of the transcriptional regulation of *NODAL* in cancer cells is far less comprehensive, it has been demonstrated that embryonic enhancers such as the NDE are active in the promotion of *NODAL* expression in response to hypoxia via induced Notch signalling in both breast cancer and melanoma cells [128].

1.14 Co-transcriptional regulation

Beyond control of transcription initiation, there are several points at which protein-coding gene expression is controlled that occur concomitant with or directly after transcription and are generally referred to as co-transcriptional regulatory mechanisms. These include, but are not limited to, transcriptional start site selection, (alternative) mRNA splicing, and the coupled processes of transcription termination and polyadenylation (reviewed in [129, 130]). For a typical transcript, these processes contribute to the identity of a transcribed and fully processed mRNA. First, the 5' end of the transcript is defined by the transcriptional start site and marks the start of the 5' untranslated region (UTR) upstream of the translational start site. The AUG methionine "start" codon marks the end of the 5' UTR and the first codon to be read by the translational machinery during translation. During transcription, removal of pre-mRNA introns and subsequent joining of flanking exons takes place in a process known as splicing. Splice donor sites define the 5' ends of

introns, while downstream splice acceptor sites define their corresponding 3' ends. Some mRNAs have only a single exon and thus do not undergo splicing, but the vast majority contain at least two exons separated by intronic sequences. A TGA, TAG, or TAA "stop" codon marks the end of the translated open reading frame [131]. The stop codon also marks the start of the 3' UTR. The 3' end of a translationally-competent mRNA is also modified to terminate in a stretch of non-templated adenine (A) bases known as the polyA tail. A corresponding modification is also made to the 5' end of messages, in the addition of a single methylated guanosine base. Collectively, these sites define the termini of transcripts (reviewed in [132]). The full-length nature of *NODAL* transcripts has not been specifically assessed.

1.15 Alternative splicing

Alternative splicing of messenger RNA is perhaps the most well-studied aspect of co-transcriptional regulation of gene expression (reviewed in [129]). Mechanisms of alternative splicing (AS) exist for many genes, whereby different combinations of exons within a single pre-mRNA can be included in distinctly processed transcripts. Alternative splicing, as for splicing in general, actually takes place co-transcriptionally [133]. According to the kinetic model, a slower rate of transcription allows inclusion of weak alternative exons, while the recruitment model posits that specific splicing factors can bind RNA polymerase II to increase their local concentration at target splice sites and thereby strengthen the interaction [134, 135].

1.16 Mechanisms of alternative splicing

The effects of growth factors and other components of the microenvironment on AS are mediated by intracellular signaling cascades [136, 137]. These cascades ultimately affect splicing regulatory proteins such as members of the serine and arginine-rich (SR) and heterogeneous nuclear ribonucleoprotein (hnRNP) families [138, 139]. Two "splicing hubs" through which a multitude of cascades converge to regulate AS have been identified as hnRNP K and Sam68 (reviewed in [139]). hnRNP K has been shown to bind pre-mRNA splicing enhancers and silencers, with direct phosphorylation of hnRNP K by Src-kinases, Protein Kinase C (PKC), ERK1/2, and JNK altering protein-protein and

protein-RNA binding patterns of this splicing factor [140]. Similarly, Sam68 binds elements within pre-mRNA and can become phosphorylated by kinases such as ERK [141]. These splicing hub proteins mediate normal developmental AS programs, and also contribute to pathology in various cancers. Expression levels of hnRNP K and Sam68 are often altered in cancer [142, 143]. Furthermore, aberrant upstream signalling in cancer such as hyperactive MAPK signalling can alter normal post-translational modifications of splicing factors, affecting their localization and function [141, 144]. Collectively, these alterations in expression and activity of splicing factors hnRNP K and Sam68 can disrupt normal splicing patterns of target transcripts important in pro-tumourigenic phenotypes [145].

For splicing to take place, splice donor and splice acceptor motifs must be precisely recognized by the splicing machinery. This machinery is referred to as the spliceosome, which consists of five nuclear ribonucleoprotein particles (snRNPs), and a large number of proteins. There are two different spliceosome complexes known as the major and minor spliceosomes that are involved in the removal of introns with different sequence features at splice donor, branch point, and splice acceptor sites. The major or U2 snRNP-dependent spliceosome catalyzes the removal of the majority (>99.5%) of introns. Of these introns, the vast majority (99%) contain GU and AG dinucleotides at their 5' and 3' ends, respectively. Notably, about 0.7% of U2 introns are defined by terminal GC and AG dinucleotides. Conversely, the minor or U12 snRNP-dependent spliceosome catalyzes the removal of less than 0.5% of human introns [146]. While it was originally thought that U12 introns universally contained AU and AC terminal dinucleotides, it was later revealed that U12 introns are instead primarily defined by specific and highly conserved sequence motifs relative to U2 introns at their 5' splice donor sites and branch point sequences, and contain both GU-AG and AU-AC terminal dinucleotides [146, 147].

Mutation of constitutive splice sites in tumour-suppressor genes may disrupt normal gene processing and thus offer a selective advantage for growth in cancer cells. For example, in breast and ovarian cancer, mutations in the tumor suppressor BRCA1 often disrupt constitutive splice sites, leading to loss of functional protein [148]. Indeed, bioinformatic tools to predict the impact of virtually any cancer-associated mutation on patterns of

alternative splicing have been developed [149-152]. For human *NODAL*, inheritance of a rare mutation within a splice site motif of a constitutively spliced exon is associated with abnormal development [153]. Notably, mutations or polymorphisms can also result in cryptic splice sites and resultant alternative splicing in regions normally constitutively spliced out as introns.

1.17 Types of alternative splicing

The relative positions of utilized 5′ donor and 3′ acceptor splice sites are used to classify different types of AS events as cassette alternative exon (or exon skipping), mutually exclusive exon, alternative 5′ splice site, alternative 3′ splice site, or complete intron retention. Cassette alternative exon splicing is the most common form of AS in humans [154]. Beyond differential inclusion of exons in processed transcripts, a more exotic form of splicing produces circular RNAs through "back-splicing" of downstream 5' splice donor sites that form junctions with upstream 3' splice donor sites of either their own exon or upstream exons, resulting in completely closed circular RNA transcripts lacking free ends. Although often generated from protein coding pre-mRNAs, these transcripts are not generally protein-coding, but can act to regulate gene expression either at the level of transcription, or post-transcriptionally through modulation of miRNA activity [155].

1.18 Widespread alternative splicing of human genes

Genome-wide analyses suggest that AS might affect as many as 95% of multi-exon human transcripts [156]. This newfound appreciation of the ubiquity of AS suggests there are numerous alternatively spliced transcript isoforms yet to be characterized.

As is true for gene transcription, patterns of AS also differ between tissues such as brain, skeletal muscle, breast, liver, and colon. Some alternatively spliced variants are virtually absent in one tissue, and constitute virtually all expressed transcripts of that gene in another tissue [156]. It follows that AS is tightly regulated over the course of development, and that specific patterns of splicing must be maintained in adult tissues to preserve distinct cellular identities and functions. Along with widespread changes in gene expression, reprogramming of AS coincides with EMT [157, 158], indicating that AS

plays an important role in phenotypic plasticity. It has been proposed that regulation of numerous alternative splicing events is part of a coordinated EMT "splicing signature." Such signatures have proven useful in the classification of breast cancer cell lines as either luminal (generally poorly metastatic) or basal (generally more aggressive and metastatic) [157]. The existence of splicing signatures for other processes integral to tumor progression such as angiogenesis has also been hypothesized [159-162]. In human embryonic stem cells, induced differentiation is accompanied by widespread changes in alternative splicing [163, 164]. A switch in the alternative splicing of a key regulator of stem cell pluripotency and differentiation also dictates cell fate [165]. Furthermore, splice variants have been described for two of the most well-studied "core" pluripotency transcription factors OCT4 [166, 167] and NANOG [168, 169]. Perhaps unsurprisingly, alternative splicing is frequently deregulated in cancer [138, 170, 171], and cancer cells can hijack stem cell alternative splicing programs to enhance the maintenance of cancer stem cells [172].

Evidently, AS is important for normal cell function, and dysregulation of AS is widespread in cancer. Therefore, AS may present opportunities for therapeutic intervention and novel prognostic biomarker identification for specific cancers [173, 174]. In addition, alternatively spliced gene products of cancer therapy targets must be extensively characterized to ensure desired targeting. Going forward, if these goals are to be achieved, alternatively spliced transcripts will need to be carefully documented, characterized, and incorporated into modelling of gene function in models of normal and malignant cell function.

1.19 Impact of alternative splicing on the human proteome

Although alternative splicing takes place at the RNA level, its manifestation at the level of corresponding translated protein products has always been of great interest. Alternative splicing is widely touted as a major contributor to the generation of proteomic diversity from a limited genome. Despite this realization, the extent to which alternative splicing contributes to productive translation of multiple protein isoforms from a single locus on a genome-wide scale remains unclear and controversial [175]. As a result, there is a distinct lack of predictive tools to decipher if a novel alternative splicing event is likely to be

biologically relevant at the protein level.

One study has attempted to identify defining features of bona fide alternative splicing events for which multiple protein isoforms have been experimentally confirmed [176], and will be briefly reviewed here: When focusing on cases where alternative splicing leads to truncation of conserved protein domains, the authors found that experimentally confirmed alternatively spliced protein isoforms always satisfied at least one of the two following criteria: truncated domain size/original domain size >0.6, or truncated domain size/protein size <0.3. That is, there were no experimentally confirmed cases where the truncated domain size/original domain size was very low, AND the truncated domain size/protein size was large. Thus, alternative splicing events leading to substantial domain truncation of large domains are unlikely to result in stable protein products. While this represents an exciting finding, such cases represented only 10% of all putative alternatively spliced variant entries in Swissprot. Therefore, identification of a typical domain disruption event where truncated domain size/original domain size >0.6, and/or truncated domain size/protein size <0.3 does not have much predictive value. Similarly, analysis of alternative splicing events validated at the RNA level revealed increased frequency of domain truncations with truncated domain size/original domain size between 90 and 100% relative to all entries of alternative splice variants. When the same comparison was made for percentage of protein disorder in the region affected by alternative splicing, validated alternative splicing events were less likely to have 0-10% disorder, and more likely to have 90-100% disorder. However, this finding was again not applicable to any individual case of alternative splicing, as alternatively spliced regions with 0-10% disorder were much more frequent than those with 90-100% disorder overall. Lastly, while this study reported 505 “minor” isoforms with evidence of expression at the protein level, a comprehensive search of the protein data bank (PDB) revealed only 15 genes for which experimentally confirmed protein structures corresponding to multiple isoforms have been obtained. This underscores the dramatic lack of genome-wide characterization of alternative splicing at the protein level.

Recently, an impressive large scale screen of protein-protein interactions for a collection of human open reading frames revealed functional significance of alternative splicing at

the proteome level on a genome wide scale [177]. This study found that linear motifs were more frequent in isoform-specific regions associated with promoting protein-protein interactions, and that their interaction partner proteins were more likely to contain linear motif binding domains than proteins involved in non-isoform-specific interactions. Alternative splicing events resulting in the truncation of conserved protein domains were also enriched for protein-protein interaction losses relative to alternative splicing events resulting in truncation in general. Quantitative analysis of protein-protein interactions for alternatively spliced proteoforms revealed cases with identical, intermediate, and completely distinct interaction profiles. Analysis also revealed that alternatively spliced proteoforms were indistinguishable from protein products of distinct genes in their interaction networks and disease associations. Isoform pairs with the most dramatic "rewiring" of protein-protein interactions were enriched for intrinsically disordered regions relative to alternatively spliced pairs with more similar interaction networks [178-180]. Another study revealed that alternatively spliced exons with tissue-specific expression were enriched for phosphorylation sites [181]. The extent to which alternative splicing modulates other post-translational modifications has not yet been assessed. Collectively, these studies suggest that alternative splicing is a bona fide mechanism for the modulation of biologically relevant protein function and interaction networks at the protein level.

1.20 The functional impact of alternative splicing

The alternative splicing of VEGFA is an excellent example of the ability of AS to confer functional divergence to products of the same gene. VEGFA is integral to angiogenesis—the expansion of blood vessel networks essential for normal tissue development and a hallmark of cancer (reviewed in [2, 3, 182]). Although VEGFA gene products are generally pro-angiogenic, alternative splicing yields a subset of VEGFA transcript isoforms that display anti-angiogenic activity (reviewed in [160]) that are in fact the predominant class of isoforms in most normal adult tissues [183]. Remarkably, these isoforms differ from their pro-angiogenic counterparts in only the six most C-terminal amino acids resulting from alternative utilization of nearby splice acceptor sites in the most 3' exon. A splicing switch promotes expression of the pro-angiogenic isoforms in

cancer [162, 184].

Despite the prevalence of alternative mRNA splicing, no alternatively spliced transcripts for the human *NODAL* gene have been described. During writing of this thesis, an alternative transcript annotation (NM_001329906.1) was added to the NCBI RefSeq database that utilizes an alternative first exon relative to the primary *NODAL* isoform (NM_018055.4). Still, no alternative splicing of *NODAL* transcripts has been described, and no putative isoforms have been characterized at either the transcript or protein levels.

1.21 Transcript cleavage and polyadenylation

Downstream of splicing events, the identity of the 3' UTR of a mature mRNA is determined by the coupled processes of pre-mRNA cleavage and polyadenylation. There are several sequence elements that guide the selection of polyadenylation sites (reviewed in [130, 185]), referred to as the upstream sequence element (USE) [186], polyadenylation signal (PAS) [187], and downstream sequence element (DSE) [188, 189]. The highly conserved PAS is found 10-30 bases upstream of the mRNA cleavage site. Analysis of PAS sequences at 7,000 bona fide human mRNA cleavage sites revealed that two motifs account for the majority of sites, with AAUAAA and AUUAAA accounting for 47% and 16% of all sites, respectively [190]. The USE is less well-defined, while the DSE is a U- or GU-rich element. As with splicing, polyadenylation can occur at multiple sites for a single transcript in a process known as alternative polyadenylation (APA) (reviewed in [130, 185]). Interestingly, different tissues show global preferences for the selection of either more distal PAS resulting in longer 3' UTRs (e.g. brain), or more proximal PAS resulting in shorter 3' UTRs (e.g. blood) [191]. Patterns of APA are also dynamic during development, with distal site selection becoming favoured during differentiation and embryonic development [192], whereas high levels of cell proliferation found in cancer and reprogramming of somatic cells to iPS cells involves selection of more proximal PAS and generally shorter 3' UTRs [193-195].

1.22 Post-transcriptional regulation of gene expression

Subsequent points of regulation are often broadly referred to as post-transcriptional

regulation. At the RNA level, these include regulation of mRNA nuclear export and stability, and regulation by complementary RNAs such as microRNAs and natural antisense transcripts (NATs). Antisense transcription occurs when two transcripts are expressed from the same genomic locus—one from each complementary strand of the genome. These transcripts are transcribed in opposite directions to yield RNAs with complementary sequence—the extent of which depends on their degree of genomic overlap. Unsurprisingly, the complementary nature of natural antisense transcripts often confers the ability of one transcript to regulate the expression (translation or otherwise) of its antisense counterpart (reviewed in [196, 197]). Although there is a putative antisense transcript in GenBank (accession AK001176) mapping to the constitutive exon 2 *NODAL* locus, this transcript has not been curated into the RefSeq database, and has not been directly studied.

Further points of post-transcriptional regulation include control over protein translation, protein trafficking and enzymatic processing, quaternary protein complex formation, and post-translational modification (PTM) of specific amino acid side chains of the protein. PTMs are integral to normal cell function and are most widely appreciated for their role in the modulation of enzyme activity through phosphorylation [198]. Several classes of PTMs play numerous roles in a myriad of cellular processes including signal transduction, protein folding and stability, and protein-protein interactions (reviewed in [199]). Unsurprisingly, PTMs also play numerous roles in the regulation of human embryonic stem cell pluripotency [200].

One post-translational modification characteristic of TGF-beta superfamily members and secreted proteins in general is N-glycosylation, which consists of the covalent addition of a glycan oligosaccharide to asparagine residues within N-X-S/T motifs [201]. N-glycosylation generally aids in protein folding in the ER, and impacts both protein secretion and stability (reviewed in [202]). As examples, extensive N-glycosylation is required for dimerization of Quercetin 2,3-dioxygenase subunits [203], and N-glycosylation of TGF betas promote the secretion of active ligand [204]. Intracellular full-length/pro-Nodal is found in an N-glycosylated form, and corresponding pro-Nodal secreted into conditioned media was found to contain complex carbohydrate

modifications, indicative of further N-glycan processing along the secretory pathway [205]. Similar modifications to both full-length pro-Nodal and the cleaved pro-domain indicate that the pro-domain is the site of these post-translational modifications. In contrast to the pro-domain, the mature peptides of both human and mouse mature Nodal ligands do not contain N-glycosylation sites. Once cleaved from the N-glycosylated pro-domain, it has been suggested that the mature Nodal peptide is rapidly degraded and thus limited in its signalling range [206]. Interestingly, experimental introduction of different N-glycosylation motifs found in BMP6 or the Xenopus nodal related (Xnr) proteins into the Nodal mature domain increased the accumulation of mature Nodal peptide in conditioned media and consequently signalling range in zebrafish blastulae [206]. However, the effect of this N-glycosylation on Nodal secretion, processing, or dimerization was not reported. Furthermore, the specific residues in the pro-domain at which endogenous N-glycosylations take place have not been directly studied, nor has the impact of these modifications on NODAL processing.

1.23 Genetics is the basis for many aspects of gene expression

One common thread to all of the processes discussed above, and indeed virtually every process in the cell, is that they are influenced by sequences in genomic DNA. Prominent examples discussed above include transcription factor binding sites, splice site dinucleotides, and polyadenylation signals. Beyond these elements, DNA obviously also templates the transcription of complementary RNA, interpreted as codons by the translational machinery, and thus the amino acid identity of cellular proteins. Many PTMs such as N-glycosylations are catalyzed at strict consensus sequences that are therefore templated by genomic DNA. Collectively, these aspects of genomic DNA underscore the impact of widespread DNA mutation on gene expression and cellular function in cancer. Even in the absence of cancer, genomic DNA is not static between generations and individuals, as non-lethal germline mutations occurring at low frequencies over evolutionary time persist in populations [207].

1.24 Genetic variation in human populations

Genetic variation between human individuals and within populations pose challenges to

biomedical research in terms of heterogeneity between individuals. Traditionally overlooked, the importance of characterizing genetic variation and considering its impact on modelling biological processes is now becoming increasingly appreciated. Going forward, these considerations will contribute to research findings that more readily translate to humans and can be incorporated into highly sought after personalized medicine approaches for combatting diseases such as cancer.

Ever since the completion of the human genome project between 2000 and 2003, there has been an intensified interest in inherited genetic variation in humans. The first step toward incorporating genetic variation into experimental models is to survey the extent and nature of genetic heterogeneity on a global scale. The 1000 Genomes Project is the most comprehensive project ever completed to catalogue this variation [207]. Recently completed in 2015, this project employed various genotyping technologies including deep sequencing to reconstruct the genomes of 2054 individuals from 26 populations representing different ancestries from around the world. The 1000 Genomes Project has detailed over 88 million genetic variants. By far the most common type of genetic variation in humans is the single nucleotide polymorphism (SNP), representing approximately 84.7 million or 96% of the variants detected [207]. A typical genome deviates from the reference genome at about 4.1 to 5.0 million sites, or about 0.15% of the genome [207]. While most variants in the entire catalogue are rare (73% have a frequency < 0.5%), most variants in a given genome are common; between 96% and 99% have a frequency of > 0.5% [207].

Although the percentage of polymorphic bases in a typical genome (0.15%) may seem underwhelming, the putative functional impact of these polymorphisms is staggering: A typical genome is estimated to contain between 149 and 182 SNP alleles resulting in protein truncation, 10,000 to 12,000 SNP alleles that alter peptide sequence, and roughly 500,000 SNPs in known regulatory regions [207].

1.25 Genome-wide association studies

There has also been a great deal of interest in identifying genetic variations or SNPs that are responsible for variation in human traits, including susceptibility to complex diseases

such as cancers, as well as response to and tolerance of specific classes of drugs. The simplest study design to identify such SNPs is to perform a genome-wide association study (GWAS) that identifies SNPs with different genotype frequencies between two populations of interest, for example subjects who have received a cancer diagnosis and subjects who have not.

1.26 The challenges and benefits of linkage disequilibrium

A major complicating aspect to identifying potentially functional SNPs from association studies is linkage disequilibrium [208]. Since any given SNP allele is generally inherited as part of an entire chromosome, it is inherited along with numerous other SNP alleles known as a haplotype. Two SNP alleles that are always inherited together throughout a population are said to be in perfect linkage disequilibrium (LD). Thus, if one of these SNPs was causally responsible for the given trait of interest and the other SNP had no function, the two SNPs would be indistinguishable in a GWAS. Although useful to reduce genotyping costs and for imputation of unknown SNPs, LD has remained a major obstacle to the identification of causal genomic variants.

Early expectations were that GWA studies would uncover numerous variants in protein-coding regions that dramatically affected protein function. Perhaps surprisingly, most GWAS hits or trait/disease-associated SNPs (TASs) instead lie in either intronic or intergenic non protein-coding regions. The NHGRI-EBI GWAS Catalog reported such variants to make up 88% of GWAS hits [209]. These high rates are retained even after more complex fine mapping approaches have been applied (e.g. 90% in [210]). These findings not only suggest that non-protein coding regions of the genome are undoubtedly functionally important, but also demand increased efforts to functionally annotate non-coding regions of the human genome.

The most massive effort to extensively functionally annotate the human genome has been the Encyclopedia Of DNA Elements (ENCODE) project. This project maps results from numerous genome-wide studies including transcription factor and histone protein ChIP-seq, DNase sensitivity assays, and RNA-seq to the human genome [211, 212]. These annotations can be extremely useful in assessing the potential function of a candidate

SNP based on its genomic location.

Many other types of data from genotyped samples are also useful for directing further study of SNPs of interest. As an example, expression quantitative trait loci (eQTL) studies link SNP alleles with expression of genes in cis or even global gene expression in trans [213], and splice site software can be used to predict how a given SNP may affect proximal splice site selection, possibly through modulation of splice site motifs [149].

1.27 SNPs in the human *NODAL* gene locus

Within the human *NODAL* gene locus, there are 630 total SNPs. Of these, 39 have a minor allele frequency (MAF) of >1% (dbSNP build 147 from UCSC Genome Browser). There are seven SNPs within the *NODAL* gene with ClinVar annotations. Of these, three are listed as "pathogenic" or "likely pathogenic." These three SNPs are associated with a developmental condition known as situs ambiguus, also known as visceral heterotaxy [153, 214], which is characterized by the random orientation of organs such as the heart, lungs, liver, spleen, and stomach, with respect to the left-right body axis (OMIM.org). The minor allele for rs104894169 results in the single amino acid change R183Q and the minor allele for rs121909283 results in the single amino acid change G260R, while the minor allele for rs878855044 results in abrogation of the constitutive exon 2 splice donor site. Beyond these annotated ClinVar polymorphisms, numerous other rare family-specific *NODAL* polymorphisms have been found in the genomes of individuals suffering from heterotaxy and other laterality abnormalities, a plethora of congenital heart defects (CHD), and holoprosencephaly (HPE)—a failure of the developing forebrain to divide into two separate hemispheres [153, 215]. Roessler and colleagues [215] used a NODAL signalling luciferase reporter in zebrafish embryos to quantify the signalling capacities of various NODAL proteins harbouring numerous polymorphisms and mutations. Many of the polymorphisms associated with abnormal developmental phenotypes conferred reduced signalling capacity upon NODAL. Interestingly, this was true of the extremely common minor allele for *NODAL* SNP rs1904589 resulting in amino acid substitution H165R, along with several other mutations in the NODAL pro-domain. Mutations in other components of the NODAL signalling pathway resulting in reduced NODAL signal strength have also been linked to HPE as well as heart and laterality defects [215, 216].

Importantly, many of these polymorphisms are very rare and therefore do not provide enough statistical power for association analysis with typical cohort sizes. To date, no GWAS associations have been described for any *NODAL* SNPs. The study by Roessler and colleagues has also been the only report to functionally assess genetic polymorphisms at the *NODAL* locus. Furthermore, only polymorphisms in coding regions were functionally assessed. This is indicative of a general inability to predict the effect of, or experimentally model, non-coding polymorphisms.

1.28 The advent of precision genome editing

Traditionally, direct functional study of SNPs has been limited to over expression of different plasmid constructs where the genetics of interest can be easily manipulated. However, such systems do not recapitulate the endogenous genomic context. Thus, it is very difficult to model non-coding polymorphisms, especially in cases where the potential functional impact of the SNP is unknown. Thankfully, recent advances in precision genome editing now potentiate the ability to modulate endogenous SNP alleles of interest in cultured human cells. Technologies such as the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/CRISPR-associated (Cas) [217]and Transcription Activator-Like Effector Nuclease (TALEN) [218] systems allow rapid construction of engineered nucleases for virtually any target of interest and have been quickly adopted by numerous fields conducting molecular biological research [219-222]. These technologies have already been used to mutate disease-associated SNP alleles (e.g. [223]). Furthermore, comprehensive computational and experimental pipelines for the mutation of SNP alleles have started to emerge [224], and were used to endogenously manipulate a cancer-associated SNP for the first time. Notably, this SNP was a non-coding intronic SNP. Precise editing of endogenous SNP alleles is the holy grail of experimental models to assess SNP function, and will unquestionably lead to the validation and/or invalidation of countless putative functional SNPs in the coming years, with tremendous implications for advancing goals of personalized medicine.

Beyond SNP editing, precision genome editing has many other applications. Perhaps the most appealing application of precision genome editing is functional gene knockout. This can be achieved by exploiting the error-prone non-homologous end joining (NHEJ)

pathway active in nuclease-induced double-stranded DNA break repair. This process results in short indel mutations at the target site [225]. Cells with translational frameshift-altering mutations in all alleles will not translate normal protein and can be used as knockout models. The use of precision nucleases also greatly enhances gene targeting abilities and introduction of exogenous constructs into the genome; a process that was previously extremely inefficient in human pluripotent stem cells despite much success in mouse counterparts (reviewed in [226]). To date, precision genome editing has not been used in any fashion to functionally knockout or otherwise study the human *NODAL* gene locus.

1.29 Thesis rationale, hypothesis, and aims

There is currently only one human *NODAL* transcript isoform that has been characterized. However, genome-wide transcriptome profiling suggests that multiple transcripts are expressed from virtually all multi-exon human genes [227]. I hypothesize that there is more than one distinct transcript expressed from the human *NODAL* locus. Comprehensive analysis will be performed to identify and characterize potential novel *NODAL* locus transcripts (chapter 3). I will also explore how genetic heterogeneity can regulate expression of novel NODAL transcripts (chapter 2), and how their translation impacts the processing and function of NODAL protein (chapter 4). Thus, I will characterize human *NODAL* gene expression at multiple levels, with an emphasis on how these levels are inter-connected. Lastly, I aim to develop tools to streamline precision genome editing workflows, and use these tools to establish robust over-expression and functional knockout *NODAL* models. Collectively, elucidation of these molecular details and development of genetic models of *NODAL* function will enrich our understanding of human-specific *NODAL* biology in models of development and disease.

1.30 References

1. Canadian Cancer Society. Cancer statistics at a glance. Retrieved October 23, 2016 from http://www.cancer.ca/en/cancer-information/cancer-101/cancer-statistics-at-a-glance/?region=on

2. Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, *144*(5), 646–674. doi:10.1016/j.cell.2011.02.013

3. Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, *100*(1), 57–70.

4. Valastyan, S., & Weinberg, R. A. (2011). Tumor Metastasis: Molecular Insights and Evolving Paradigms. *Cell*, *147*(2), 275–292. doi:10.1016/j.cell.2011.09.024

5. Holohan, C., Van Schaeybroeck, S., Longley, D. B., & Johnston, P. G. (2013). Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer*, *13*(10), 714–726. doi:10.1038/nrc3599

6. Dick, J. E. (2009, January). Looking ahead in cancer stem cell research. *Nature Biotechnology*, pp. 44–46. doi:10.1038/nbt0109-44

7. Rowan, K. (2009). Are Cancer Stem Cells Real? After Four Decades, Debate Still Simmers. *JNCI Journal of the National Cancer Institute*, *101*(8), 546–547. doi:10.1093/jnci/djp083

8. Jordan, C. T. (2009). Cancer Stem Cells: Controversial or Just Misunderstood? *Cell Stem Cell*, *4*(3), 203–205. doi:10.1016/j.stem.2009.02.003

9. Enderling, H. (2015). Cancer stem cells: small subpopulation or evolving fraction? *Integr. Biol.*, *7*(1), 14–23. doi:10.1039/C4IB00191E

10. Takebe, N., & Ivy, S. P. (2010). Controversies in Cancer Stem Cells: Targeting Embryonic Signaling Pathways. *Clinical Cancer Research*, *16*(12), 3106–3112. doi:10.1158/1078-0432.CCR-09-2934

11. Krock, B. L., Skuli, N., & Simon, M. C. (2011). Hypoxia-induced angiogenesis: good and evil. *Genes & cancer*, *2*(12), 1117–1133. doi:10.1177/1947601911423654

12. Easwaran, H., Tsai, H.-C., & Baylin, S. B. (2014). Cancer Epigenetics: Tumor Heterogeneity, Plasticity of Stem-like States, and Drug Resistance. *Molecular Cell*, *54*(5), 716–727. doi:10.1016/j.molcel.2014.05.015

13. Kalluri, R., & Weinberg, R. A. (2009). The basics of epithelial-mesenchymal transition. *Journal of Clinical Investigation*, *119*(6), 1420–1428. doi:10.1172/JCI39104

14. Hay, E. D. (1995). An overview of epithelio-mesenchymal transformation. *Acta anatomica*, *154*(1), 8–20.

15. Lee, J. M., Dedhar, S., Kalluri, R., & Thompson, E. W. (2006). The epithelial–mesenchymal transition: new insights in signaling, development, and disease. *The Journal of Cell Biology*, *172*(7), 973–981. doi:10.1083/jcb.200601018

16. van Denderen, B. J. W., & Thompson, E. W. (2013, January 24). Cancer: The to and fro of tumour spread. *Nature*, pp. 487–488. doi:10.1038/493487a

17. Thomson, J. A., Itskovitz-Eldor, J., Shapiro, S. S., Waknitz, M. A., Swiergiel, J. J., Marshall, V. S., & Jones, J. M. (1998). Embryonic stem cell lines derived from human blastocysts. *Science*, *282*(5391), 1145–1147.

18. Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., & Yamanaka, S. (2007). Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell*, *131*(5), 861–872. doi:10.1016/j.cell.2007.11.019

19. Kim. (2015). Applications of iPSCs in Cancer Research. *Biomarker Insights*, 125–7. doi:10.4137/BMI.S20065

20. O'Leary, T., Heindryckx, B., Lierman, S., van Bruggen, D., Goeman, J. J., Vandewoestyne, M., et al. (2012). Tracking the progression of the human inner cell mass during embryonic stem cell derivation. *Nature Biotechnology*, *30*(3), 278–282. doi:10.1038/nbt.2135

21. Brons, I. G. M., Smithers, L. E., Trotter, M. W. B., Rugg-Gunn, P., Sun, B., Chuva de Sousa Lopes, S. M., et al. (2007). Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature*, *448*(7150), 191–195. doi:10.1038/nature05950

22. Tesar, P. J., Chenoweth, J. G., Brook, F. A., Davies, T. J., Evans, E. P., Mack, D. L., et al. (2007). New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature*, *448*(7150), 196–199. doi:10.1038/nature05972

23. Vallier, L., Touboul, T., Chng, Z., Brimpari, M., Hannan, N., Millan, E., et al. (2009). Early Cell Fate Decisions of Human Embryonic Stem Cells and Mouse Epiblast Stem Cells Are Controlled by the Same Signalling Pathways. *PLoS ONE*, *4*(6), e6082. doi:10.1371/journal.pone.0006082.t002

24. Greber, B., Wu, G., Bernemann, C., Joo, J. Y., Han, D. W., Ko, K., et al. (2010). Conserved and Divergent Roles of FGF Signaling in Mouse Epiblast Stem Cells and Human Embryonic Stem Cells. *Cell Stem Cell*, *6*(3), 215–226. doi:10.1016/j.stem.2010.01.003

25. Pan, G., Tian, S., Nie, J., Yang, C., Ruotti, V., Wei, H., et al. (2007). Whole-Genome Analysis of Histone H3 Lysine 4 and Lysine 27 Methylation in Human Embryonic Stem Cells. *Cell Stem Cell*, *1*(3), 299–312. doi:10.1016/j.stem.2007.08.003

26. Delgado-Olguin, P., & Recillas-Targa, F. (2011). Chromatin structure of pluripotent stem cells and induced pluripotent stem cells. *Briefings in Functional Genomics*, *10*(1), 37–49. doi:10.1093/bfgp/elq038

27. Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., et

al. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, *125*(2), 315–326. doi:10.1016/j.cell.2006.02.041

28. Hanna, J., Cheng, A. W., Saha, K., Kim, J., Lengner, C. J., Soldner, F., et al. (2010). Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(20), 9222–9227. doi:10.1073/pnas.1004584107

29. Gafni, O., Weinberger, L., Mansour, A. A., Manor, Y. S., Chomsky, E., Ben-Yosef, D., et al. (2013). Derivation of novel human ground state naive pluripotent stem cells. *Nature*, *504*(7479), 282–286. doi:10.1038/nature12745

30. Wu, J., & Izpisua Belmonte, J. C. (2015). Dynamic Pluripotent Stem Cell States and Their Applications. *Cell Stem Cell*, *17*(5), 509–525. doi:10.1016/j.stem.2015.10.009

31. Bertero, A., Madrigal, P., Galli, A., Hubner, N. C., Moreno, I., Burks, D., et al. (2015). Activin/Nodal signaling and NANOG orchestrate human embryonic stem cell fate decisions by controlling the H3K4me3 chromatin mark. *Genes & Development*, *29*(7), 702–717. doi:10.1101/gad.255984.114

32. Xi, Q., Wang, Z., Zaromytidou, A.-I., Zhang, X. H. F., Chow-Tsang, L.-F., Liu, J. X., et al. (2011). A poised chromatin platform for TGF-β access to master regulators. *Cell*, *147*(7), 1511–1524. doi:10.1016/j.cell.2011.11.032

33. Kim, S. W., Yoon, S.-J., Chuong, E., Oyolu, C., Wills, A. E., Gupta, R., & Baker, J. (2011). Chromatin and transcriptional signatures for Nodal signaling during endoderm formation in hESCs. *Developmental Biology*, *357*(2), 492–504. doi:10.1016/j.ydbio.2011.06.009

34. Brown, S., Teo, A., Pauklin, S., Hannan, N., Cho, C. H. H., Lim, B., et al. (2011). Activin/Nodal Signaling Controls Divergent Transcriptional Networks in Human Embryonic Stem Cells and in Endoderm Progenitors. *STEM CELLS*, *29*(8), 1176–1185. doi:10.1002/stem.666

35. Weiss, A., & Attisano, L. (2012). The TGFbeta Superfamily Signaling Pathway. *Wiley Interdisciplinary Reviews: Developmental Biology*, *2*(1), 47–63. doi:10.1002/wdev.86

36. Iyer, S., & Acharya, K. R. (2011). Tying the knot: The cystine signature and molecular-recognition processes of the vascular endothelial growth factor family of angiogenic cytokines. *FEBS Journal*, *278*(22), 4304–4322. doi:10.1111/j.1742-4658.2011.08350.x

37. Wrana, J. L., Attisano, L., Wieser, R., Ventura, F., & Massague, J. (1994). Mechanism of activation of the TGF-beta receptor. *Nature*.

38. Wrana, J. L., Attisano, L., Cárcamo, J., Zentella, A., Doody, J., Laiho, M., et al. (1992). TGFβ signals through a heteromeric protein kinase receptor complex. *Cell*, *71*(6), 1003–1014. doi:10.1016/0092-8674(92)90395-S

39. Gritsman, K., Zhang, J., Cheng, S., Heckscher, E., Talbot, W. S., & Schier, A. F. (1999). The EGF-CFC Protein One-Eyed Pinhead Is Essential for Nodal Signaling. *Cell*, *97*(1), 121–132. doi:10.1016/S0092-8674(00)80720-5

40. Constam, D. B. (2009). Riding Shotgun: A Dual Role for the Epidermal Growth Factor-Cripto/FRL-1/Cryptic Protein Cripto in Nodal Trafficking. *Traffic*, *10*(7), 783–791. doi:10.1111/j.1600-0854.2009.00874.x

41. Souchelnytskyi, S., Rönnstrand, L., Heldin, C. H., & Dijke, ten, P. (2001). Phosphorylation of Smad signaling proteins by receptor serine/threonine kinases. *Methods in molecular biology (Clifton, N.J.)*, *124*, 107–120.

42. Feng, X.-H., & Derynck, R. (2005). Specificity and versatility in tgf-beta signaling through Smads. *Annual review of cell and developmental biology*, *21*(1), 659–693. doi:10.1146/annurev.cellbio.21.022404.142018

43. Zawel, L., Dai, J. L., Buckhaults, P., Zhou, S., Kinzler, K. W., Vogelstein, B., & Kern, S. E. (1998). Human Smad3 and Smad4 are sequence-specific transcription activators. *Molecular Cell*, *1*(4), 611–617.

44. Labbé, E., Silvestri, C., Hoodless, P. A., Wrana, J. L., & Attisano, L. (1998). Smad2 and Smad3 Positively and Negatively Regulate TGFβ-Dependent Transcription through the Forkhead DNA-Binding Protein FAST2. *Molecular Cell*, *2*(1), 109–120.

45. Chen, X., Rubock, M. J., & Whitman, M. (1996). A transcriptional partner for MAD proteins in TGF-beta signalling. *Nature*, *383*(6602), 691–696. doi:10.1038/383691a0

46. Derynck, R., & Zhang, Y. E. (2003). Smad-dependent and Smad-independent pathways in TGF-beta family signalling. *Nature*, *425*(6958), 577–584. doi:10.1038/nature02006

47. Zhang, Y. E. (2009). Non-Smad pathways in TGF-|[beta]| signaling. *Cell Research*, *19*(1), 128–139. doi:10.1038/cr.2008.328

48. Guo, X., & Wang, X.-F. (2009). Signaling cross-talk between TGF-beta/BMP and other pathways. *Cell Research*, *19*(1), 71–88. doi:10.1038/cr.2008.302

49. Zhou, X., Sasaki, H., Lowe, L., Hogan, B. L., & Kuehn, M. R. (1993). Nodal is a novel TGF-beta-like gene expressed in the mouse node during gastrulation. *Nature*, *361*(6412), 543–547. doi:10.1038/361543a0

50. Davidson, B. P., & Tam, P. P. L. (2000). The node of the mouse embryo.

Current Biology, *10*(17), R617–R619. doi:10.1016/S0960-9822(00)00675-8

51. Shen, M. M. (2007). Nodal signaling: developmental roles and regulation. *Development*, *134*(6), 1023–1034. doi:10.1242/dev.000166

52. Schier, A. F. (2009). Nodal Morphogens. *Cold Spring Harbor Perspectives in Biology*, *1*(5), a003459–a003459. doi:10.1101/cshperspect.a003459

53. Quail, D. F., Siegers, G. M., Jewer, M., & Postovit, L.-M. (2013). Nodal signalling in embryogenesis and tumourigenesis. *The International Journal of Biochemistry & Cell Biology*, *45*(4), 885–898. doi:10.1016/j.biocel.2012.12.021

54. Bodenstine, T. M., Chandler, G. S., Seftor, R. E. B., Seftor, E. A., & Hendrix, M. J. C. (2016). Plasticity underlies tumor progression: role of Nodal signaling. *Cancer and Metastasis Reviews*, *35*(1), 21–39. doi:10.1007/s10555-016-9605-5

55. Beck, S., Le Good, J. A., Guzman, M., Ben-Haim, N., Roy, K., Beermann, F., & Constam, D. B. (2002). Extraembryonic proteases regulate Nodal signalling during gastrulation. *Nature Cell Biology*, *4*(12), 981–985. doi:10.1038/ncb890

56. Shen, M. M., & Schier, A. F. (2000). The EGF-CFC gene family in vertebrate development. *Trends in genetics : TIG*, *16*(7), 303–309.

57. Liguori, G. L., Borges, A. C., D'Andrea, D., Liguoro, A., Gonçalves, L., Salgueiro, A. M., et al. (2008). Cripto-independent Nodal signaling promotes positioning of the A-P axis in the early mouse embryo. *Developmental Biology*, *315*(2), 280–289. doi:10.1016/j.ydbio.2007.12.027

58. Yeo, C., & Whitman, M. (2001). Nodal signals to Smads through Cripto-dependent and Cripto-independent mechanisms. *Molecular Cell*, *7*(5), 949–957.

59. Schier, A. F., & Shen, M. M. (2000). Nodal signalling in vertebrate development. *Nature*, *403*(6768), 385–389. doi:10.1038/35000126

60. Norris, D. P., Brennan, J., Bikoff, E. K., & Robertson, E. J. (2002). The Foxh1-dependent autoregulatory enhancer controls the level of Nodal signals in the mouse embryo. *Development*, *129*(14), 3455–3468.

61. Chen, C., & Shen, M. M. (2004). Two Modes by which Lefty Proteins Inhibit Nodal Signaling. *Current Biology*, *14*(7), 618–624. doi:10.1016/j.cub.2004.02.042

62. Robertson, E. J. (2014). Dose-dependent Nodal/Smad signals pattern the early mouse embryo. *Seminars in Cell & Developmental Biology*, *32*, 73–79. doi:10.1016/j.semcdb.2014.03.028

63. Müller, P., Rogers, K. W., Jordan, B. M., Lee, J. S., Robson, D., Ramanathan, S., & Schier, A. F. (2012). Differential diffusivity of Nodal and Lefty underlies a reaction-diffusion patterning system. *Science*, *336*(6082), 721–724.

doi:10.1126/science.1221920

64. Sakuma, R., Ohnishi Yi, Y.-I., Meno, C., Fujii, H., Juan, H., Takeuchi, J., et al. (2002). Inhibition of Nodal signalling by Lefty mediated through interaction with common receptors and efficient diffusion. *Genes to cells : devoted to molecular & cellular mechanisms*, *7*(4), 401–412.

65. Juan, H., & Hamada, H. (2001). Roles of nodal-lefty regulatory loops in embryonic patterning of vertebrates. *Genes to cells : devoted to molecular & cellular mechanisms*, *6*(11), 923–930.

66. van Boxtel, A. L., Chesebro, J. E., Heliot, C., Ramel, M.-C., Stone, R. K., & Hill, C. S. (2015). A Temporal Window for Signal Activation Dictates the Dimensions of a Nodal Signaling Domain. *Developmental Cell*, *35*(2), 175–185. doi:10.1016/j.devcel.2015.09.014

67. Piccolo, S., Agius, E., Leyns, L., Bhattacharyya, S., Grunz, H., Bouwmeester, T., & De Robertis, E. M. (1999). The head inducer Cerberus is a multifunctional antagonist of Nodal, BMP and Wnt signals. *Nature*, *397*(6721), 707–710. doi:10.1038/17820

68. Camus, A., Perea-Gomez, A., Moreau, A., & Collignon, J. (2006). Absence of Nodal signaling promotes precocious neural differentiation in the mouse embryo. *Developmental Biology*, *295*(2), 743–755. doi:10.1016/j.ydbio.2006.03.047

69. Mesnard, D. (2006). Nodal specifies embryonic visceral endoderm and sustains pluripotent cells in the epiblast before overt axial patterning. *Development*, 1–9. doi:10.1242/dev.02413

70. Yamamoto, M., Saijoh, Y., Perea-Gomez, A., Shawlot, W., Behringer, R. R., Ang, S.-L., et al. (2004). Nodal antagonists regulate formation of the anteroposterior axis of the mouse embryo. *Nature*, *428*(6981), 387–392. doi:10.1038/nature02418

71. Kumar, A., Lualdi, M., Lyozin, G. T., Sharma, P., Loncarek, J., Fu, X.-Y., & Kuehn, M. R. (2015). Nodal signaling from the visceral endoderm is required to maintain Nodal gene expression in the epiblast and drive DVE/AVE migration. *Developmental Biology*, *400*(1), 1–9. doi:10.1016/j.ydbio.2014.12.016

72. Ben-Haim, N., Lu, C., Guzman-Ayala, M., Pescatore, L., Mesnard, D., Bischofberger, M., et al. (2006). The Nodal Precursor Acting via Activin Receptors Induces Mesoderm by Maintaining a Source of Its Convertases and BMP4. *Developmental Cell*, *11*(3), 313–323. doi:10.1016/j.devcel.2006.07.005

73. Conlon, F. L., Lyons, K. M., Takaesu, N., Barth, K. S., Kispert, A., Herrmann, B., & Robertson, E. J. (1994). A primary requirement for nodal in the formation and maintenance of the primitive streak in the mouse. *Development*, *120*(7),

1919–1928.

74. Brennan, J., Norris, D. P., & Robertson, E. J. (2002). Nodal activity in the node governs left-right asymmetry. *Genes & Development*, *16*(18), 2339–2344. doi:10.1101/gad.1016202

75. Mercola, M. (2003). Left-right asymmetry: nodal points. *Journal of Cell Science*, *116*(Pt 16), 3251–3257. doi:10.1242/jcs.00668

76. Adachi, H., Saijoh, Y., Mochida, K., Ohishi, S., Hashiguchi, H., Hirao, A., & Hamada, H. (1999). Determination of left/right asymmetric expression of nodal by a left side-specific enhancer with sequence similarity to a lefty-2 enhancer. *Genes & Development*, *13*(12), 1589–1600.

77. Yamamoto, M., Mine, N., Mochida, K., Sakai, Y., Saijoh, Y., Meno, C., & Hamada, H. (2003). Nodal signaling induces the midline barrier by activating Nodal expression in the lateral plate. *Development*, *130*(9), 1795–1804. doi:10.1242/dev.00408

78. Varlet, I., Collignon, J., & Robertson, E. J. (1997). nodal expression in the primitive endoderm is required for specification of the anterior axis during mouse gastrulation. *Development*, *124*(5), 1033–1044.

79. Collignon, J., Varlet, I., & Robertson, E. J. (1996). Relationship between asymmetric nodal expression and the direction of embryonic turning. *Nature*, *381*(6578), 155–158. doi:10.1038/381155a0

80. Osada, S. I., Saijoh, Y., Frisch, A., Yeo, C. Y., Adachi, H., Watanabe, M., et al. (2000). Activin/nodal responsiveness and asymmetric expression of a Xenopus nodal-related gene converge on a FAST-regulated module in intron 1. *Development*, *127*(11), 2503–2514.

81. Lowe, L. A., Yamada, S., & Kuehn, M. R. (2001). Genetic dissection of nodal function in patterning the mouse embryo. *Development*, *128*(10), 1831–1843.

82. Vallier, L. (2005). Activin/Nodal and FGF pathways cooperate to maintain pluripotency of human embryonic stem cells. *Journal of Cell Science*, *118*(19), 4495–4509. doi:10.1242/jcs.02553

83. Vallier, L., Reynolds, D., & Pedersen, R. A. (2004). Nodal inhibits differentiation of human embryonic stem cells along the neuroectodermal default pathway. *Developmental Biology*, *275*(2), 403–421. doi:10.1016/j.ydbio.2004.08.031

84. Vallier, L., Mendjan, S., Brown, S., Chng, Z., Teo, A., Smithers, L. E., et al. (2009). Activin/Nodal signalling maintains pluripotency by controlling Nanog expression. *Development*, *136*(8), 1339–1349. doi:10.1242/dev.033951

85. Topczewska, J. M., Postovit, L.-M., Margaryan, N. V., Sam, A., Hess, A. R., Wheaton, W. W., et al. (2006). Embryonic and tumorigenic pathways converge via Nodal signaling: role in melanoma aggressiveness. *Nature Medicine*, *12*(8), 925–932. doi:10.1038/nm1448

86. Meyer, M. J., Fleming, J. M., Ali, M. A., Pesesky, M. W., Ginsburg, E., & Vonderhaar, B. K. (2009). Dynamic regulation of CD24 and the invasive, CD44posCD24neg phenotype in breast cancer cell lines. *Breast cancer research : BCR*, *11*(6), R82. doi:10.1186/bcr2449

87. Quail, D. F., Zhang, G., Walsh, L. A., Siegers, G. M., Dieters-Castator, D. Z., Findlay, S. D., et al. (2012). Embryonic Morphogen Nodal Promotes Breast Cancer Growth and Progression. *PLoS ONE*, *7*(11), e48237–12. doi:10.1371/journal.pone.0048237

88. Quail, D. F., Walsh, L. A., Zhang, G., Findlay, S. D., Moreno, J., Fung, L., et al. (2012). Embryonic protein nodal promotes breast cancer vascularization. *Cancer Research*, *72*(15), 3851–3863. doi:10.1158/0008-5472.CAN-11-3951

89. Quail, D. F., Zhang, G., Findlay, S. D., Hess, D. A., & Postovit, L. M. (2013). Nodal promotes invasive phenotypes via a mitogen-activated protein kinase-dependent pathway. *Oncogene*. doi:10.1038/onc.2012.608

90. Kirsammer, G., Strizzi, L., Margaryan, N. V., Gilgur, A., Hyser, M., Atkinson, J., et al. (2014). Nodal signaling promotes a tumorigenic phenotype in human breast cancer. *Seminars in Cancer Biology*, *29*, 40–50. doi:10.1016/j.semcancer.2014.07.007

91. Lawrence, M. G., Margaryan, N. V., Loessner, D., Collins, A., Kerr, K. M., Turner, M., et al. (2011). Reactivation of embryonic nodal signaling is associated with tumor progression and promotes the growth of prostate cancer cells. *The Prostate*, *71*(11), 1198–1209. doi:10.1002/pros.21335

92. Vo, B. T., & Khan, S. A. (2011). Expression of nodal and nodal receptors in prostate stem cells and prostate cancer cells: autocrine effects on cell proliferation and migration. *The Prostate*, *71*(10), 1084–1096. doi:10.1002/pros.21326

93. Xu, G., Zhong, Y., Munir, S., Yang, B. B., Tsang, B. K., & Peng, C. (2004). Nodal Induces Apoptosis and Inhibits Proliferation in Human Epithelial Ovarian Cancer Cells via Activin Receptor-Like Kinase 7. *The Journal of Clinical Endocrinology & Metabolism*, *89*(11), 5523–5534. doi:10.1210/jc.2004-0893

94. Fu, G., & Peng, C. (2011). Nodal enhances the activity of FoxO3a and its synergistic interaction with Smads to regulate cyclin G2 transcription in ovarian cancer cells. *Oncogene*, *30*(37), 3953–3966. doi:10.1038/onc.2011.127

95. Lonardo, E., Hermann, P. C., Mueller, M.-T., Huber, S., Balic, A., Miranda-Lorenzo, I., et al. (2011). Nodal/Activin Signaling Drives Self-Renewal and Tumorigenicity of Pancreatic Cancer Stem Cells and Provides a Target for Combined Drug Therapy. *Cell Stem Cell*, *9*(5), 433–446. doi:10.1016/j.stem.2011.10.001

96. Hueng, D.-Y., Lin, G.-J., Huang, S.-H., Liu, L.-W., Ju, D.-T., Chen, Y.-W., et al. (2011). Inhibition of Nodal suppresses angiogenesis and growth of human gliomas. *Journal of neuro-oncology*, *104*(1), 21–31. doi:10.1007/s11060-010-0467-3

97. Lee, C.-C., Jan, H.-J., Lai, J.-H., Ma, H.-I., Hueng, D.-Y., Lee, Y.-C. G., et al. (2010). Nodal promotes growth and invasion in human gliomas. *Oncogene*, *29*(21), 3110–3123. doi:10.1038/onc.2010.55

98. De Silva, T., Ye, G., Liang, Y.-Y., Fu, G., Xu, G., & Peng, C. (2012). Nodal promotes glioblastoma cell growth. *Frontiers in endocrinology*, *3*, 59. doi:10.3389/fendo.2012.00059

99. Papageorgiou, I., Nicholls, P. K., Wang, F., Lackmann, M., Makanji, Y., Salamonsen, L. A., et al. (2009). Expression of nodal signalling components in cycling human endometrium and in endometrial cancer. *Reproductive Biology and Endocrinology*, *7*(1), 122–11. doi:10.1186/1477-7827-7-122

100. Cavallari, C., Fonsato, V., Herrera, M. B., Bruno, S., Tetta, C., & Camussi, G. (2013). Role of Lefty in the anti tumor activity of human adult liver stem cells. *Oncogene*, *32*(7), 819–826. doi:10.1038/onc.2012.114

101. Munir, S., Xu, G., Wu, Y., Yang, B., Lala, P. K., & Peng, C. (2004). Nodal and ALK7 inhibit proliferation and induce apoptosis in human trophoblast cells. *Journal of Biological Chemistry*, *279*(30), 31277–31286. doi:10.1074/jbc.M400641200

102. Law, J., Zhang, G., Dragan, M., Postovit, L.-M., & Bhattacharya, M. (2014). Nodal signals via β-arrestins and RalGTPases to regulate trophoblast invasion. *Cellular signalling*, *26*(9), 1935–1942. doi:10.1016/j.cellsig.2014.05.009

103. Zhong, Y., Xu, G., Ye, G., Lee, D., Modica-Amore, J., & Peng, C. (2009). Nodal and activin receptor-like kinase 7 induce apoptosis in human breast cancer cell lines: Role of caspase 3. *International journal of physiology, pathophysiology and pharmacology*, *1*(1), 83–96.

104. Strizzi, L., Hardy, K. M., Margaryan, N. V., Hillman, D. W., Seftor, E. A., Chen, B., et al. (2012). Potential for the embryonic morphogen Nodal as a prognostic and predictive biomarker in breast cancer. *Breast cancer research : BCR*, *14*(3), R75. doi:10.1186/bcr3185

105. Ning, F., Wang, H.-F., Guo, Q., Liu, Z.-C., Li, Z.-Q., & Du, J. (2015). Expression and significance of Nodal in human cancers: a meta-analysis. *International journal of clinical and experimental medicine*, *8*(11), 20227–20235.

106. Dahle, Ø., Kumar, A., & Kuehn, M. R. (2010). Nodal signaling recruits the histone demethylase Jmjd3 to counteract polycomb-mediated repression at target genes. *Science Signaling*, *3*(127), ra48–ra48. doi:10.1126/scisignal.2000841

107. Postovit, L.-M., Margaryan, N. V., Seftor, E. A., Kirschmann, D. A., Lipavsky, A., Wheaton, W. W., et al. (2008). Human embryonic stem cell microenvironment suppresses the tumorigenic phenotype of aggressive cancer cells. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(11), 4329–4334. doi:10.1073/pnas.0800467105

108. Khalkhali-Ellis, Z., Kirschmann, D. A., Seftor, E. A., Gilgur, A., Bodenstine, T. M., Hinck, A. P., & Hendrix, M. J. C. (2014). Divergence(s) in nodal signaling between aggressive melanoma and embryonic stem cells. *International Journal of Cancer*, *136*(5), E242–E251. doi:10.1002/ijc.29198

109. Kelly, R. K., Olson, D. L., Sun, Y., Wen, D., Wortham, K. A., Antognetti, G., et al. (2011). An antibody-cytotoxic conjugate, BIIB015, is a new targeted therapy for Cripto positive tumours. *European journal of cancer (Oxford, England : 1990)*, *47*(11), 1736–1746. doi:10.1016/j.ejca.2011.02.023

110. Herbertz, S., Sawyer, J. S., Stauber, A. J., Gueorguieva, I., Driscoll, K. E., Estrem, S. T., et al. (2015). Clinical development of galunisertib (LY2157299 monohydrate), a small molecule inhibitor of transforming growth factor-beta signaling pathway. *Drug design, development and therapy*, *9*, 4479–4499. doi:10.2147/DDDT.S86621

111. Focà, A., Sanguigno, L., Focà, G., Strizzi, L., Iannitti, R., Palumbo, R., et al. (2015). New Anti-Nodal Monoclonal Antibodies Targeting the Nodal Pre-Helix Loop Involved in Cripto-1 Binding. *International Journal of Molecular Sciences*, *16*(9), 21342–21362. doi:10.3390/ijms160921342

112. Strizzi, L., Sandomenico, A., Margaryan, N. V., Focà, A., Sanguigno, L., Bodenstine, T. M., et al. (2015). Effects of a novel Nodal-targeting monoclonal antibody in melanoma. *Oncotarget*, *6*(33), 34071–34086. doi:10.18632/oncotarget.6049

113. Blakeley, P., Fogarty, N. M. E., del Valle, I., Wamaitha, S. E., Hu, T. X., Elder, K., et al. (2015). Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development*, *142*(20), 3613–3613. doi:10.1242/dev.131235

114. Inman, G. J., Nicolás, F. J., Callahan, J. F., Harling, J. D., Gaster, L. M., Reith, A. D., et al. (2002). SB-431542 is a potent and specific inhibitor of transforming growth factor-beta superfamily type I activin receptor-like kinase (ALK) receptors ALK4, ALK5, and ALK7. *Molecular pharmacology*, *62*(1), 65–74.

115. James, D. (2005). TGF /activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. *Development*, *132*(6), 1273–1282. doi:10.1242/dev.01706

116. Cedar, H., & Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, *10*(5), 295–304. doi:10.1038/nrg2540

117. Sampath, K., & Robertson, E. J. (2016). Keeping a lid on nodal: transcriptional and translational repression of nodal signalling. *Open Biology*, *6*(1), 150200–8. doi:10.1098/rsob.150200

118. Norris, D. P., & Robertson, E. J. (1999). Asymmetric and node-specific nodal expression patterns are controlled by two distinct cis-acting regulatory elements. *Genes & Development*, *13*(12), 1575–1588.

119. Krebs, L. T. (2003). Notch signaling regulates left-right asymmetry determination by inducing Nodal expression. *Genes & Development*, *17*(10), 1207–1212. doi:10.1101/gad.1084703

120. Raya, A. (2003). Notch activity induces Nodal expression and mediates the establishment of left-right asymmetry in vertebrate embryos. *Genes & Development*, *17*(10), 1213–1218. doi:10.1101/gad.1084403

121. Vincent, S. D., Norris, D. P., Ann Le Good, J., Constam, D. B., & Robertson, E. J. (2004). Asymmetric Nodal expression in the mouse is governed by the combinatorial activities of two distinct regulatory elements. *Mechanisms of Development*, *121*(11), 1403–1415. doi:10.1016/j.mod.2004.06.002

122. Saijoh, Y., Oki, S., Tanaka, C., Nakamura, T., Adachi, H., Yan, Y.-T., et al. (2005). Two nodal-responsive enhancers control left-right asymmetric expression ofNodal. *Developmental Dynamics*, *232*(4), 1031–1036. doi:10.1002/dvdy.20192

123. Papanayotou, C., Benhaddou, A., Camus, A., Perea-Gomez, A., Jouneau, A., Mezger, V., et al. (2014). A Novel Nodal Enhancer Dependent on Pluripotency Factors and Smad2/3 Signaling Conditions a Regulatory Switch During Epiblast Maturation. *PLoS Biology*, *12*(6), e1001890–14. doi:10.1371/journal.pbio.1001890

124. Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., et al. (2008). Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell*, *133*(6), 1106–1117.

doi:10.1016/j.cell.2008.04.043

125. Kim, J., Chu, J., Shen, X., Wang, J., & Orkin, S. H. (2008). An Extended Transcriptional Network for Pluripotency of Embryonic Stem Cells. *Cell*, *132*(6), 1049–1061. doi:10.1016/j.cell.2008.02.039

126. Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, *122*(6), 947–956. doi:10.1016/j.cell.2005.08.020

127. Arai, D., Hayakawa, K., Ohgane, J., Hirosawa, M., Nakao, Y., Tanaka, S., & Shiota, K. (2015). An epigenetic regulatory element of the Nodal gene in the mouse and human genomes. *Mechanisms of Development*, *136*, 143–154. doi:10.1016/j.mod.2014.12.003

128. Quail, D. F., Taylor, M. J., Walsh, L. A., Dieters-Castator, D., Das, P., Jewer, M., et al. (2011). Low oxygen levels induce the expression of the embryonic morphogen Nodal. *Molecular biology of the cell*, *22*(24), 4809–4821. doi:10.1091/mbc.E11-03-0263

129. Bentley, D. L. (2014). Coupling mRNA processing with transcription in time and space. *Nature Reviews Genetics*, *15*(3), 163–175. doi:10.1038/nrg3662

130. de Klerk, E., & t Hoen, P. A. C. (2015). Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends in Genetics*, *31*(3), 128–139. doi:10.1016/j.tig.2015.01.001

131. Griffiths, A. J., Miller, J. H., Suzuki, D. T., Lewontin, R. C., & Gelbart, W. M. (2000). *An Introduction to Genetic Analysis* (7 ed.). New York: W. H. Freeman. doi:10.2307/4445755

132. Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002). Untranslated regions of mRNAs. *Genome biology*, *3*(3), REVIEWS0004.

133. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., et al. (2011). Global quantification of mammalian gene expression control. *Nature*, *473*(7347), 337–342. doi:10.1038/nature10098

134. Kornblihtt, A. R. (2007). Coupling Transcription and Alternative Splicing. In *Alternative Splicing in the Postgenomic Era* (Vol. 623, pp. 175–189). New York, NY: Springer New York. doi:10.1007/978-0-387-77374-2_11

135. Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R., & Misteli, T. (2011). Epigenetics in Alternative Pre-mRNA Splicing. *Cell*, *144*(1), 16–26. doi:10.1016/j.cell.2010.11.056

136. Li, C. Y., Chu, J. Y., Yu, J. K., Huang, X. Q., Liu, X. J., Shi, L., et al. (2004). Regulation of alternative splicing of Bcl-x by IL-6, GM-CSF and TPA. *Cell*

Research, *14*(6), 473–479. doi:10.1038/sj.cr.7290250

137. Stamm, S. (2002). Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome. *Human Molecular Genetics*, *11*(20), 2409–2416.

138. David, C. J., & Manley, J. L. (2010). Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes & Development*, *24*(21), 2343–2364. doi:10.1101/gad.1973010

139. Lynch, K. W. (2007). Regulation of alternative splicing by signal transduction pathways. *Advances in experimental medicine and biology*, *623*, 161–174.

140. Bomsztyk, K., Denisenko, O., & Ostrowski, J. (2004). hnRNP K: one protein multiple processes. *BioEssays : news and reviews in molecular, cellular and developmental biology*, *26*(6), 629–638. doi:10.1002/bies.20048

141. Matter, N., Herrlich, P., & König, H. (2002). Signal-dependent regulation of splicing via phosphorylation of Sam68. *Nature*, *420*(6916), 691–695. doi:10.1038/nature01153

142. Hope, N. R., & Murray, G. I. (2011). The expression profile of RNA-binding proteins in primary and metastatic colorectal cancer: relationship of heterogeneous nuclear ribonucleoproteins with prognosis☆. *Human Pathology*, *42*(3), 393–402. doi:10.1016/j.humpath.2010.08.006

143. Wen, F., Shen, A., Shanas, R., Bhattacharyya, A., Lian, F., Hostetter, G., & Shi, J. (2010). Higher expression of the heterogeneous nuclear ribonucleoprotein k in melanoma. *Annals of surgical oncology*, *17*(10), 2619–2627. doi:10.1245/s10434-010-1121-1

144. Lewis, T. S., Hunt, J. B., Aveline, L. D., Jonscher, K. R., Louie, D. F., Yeh, J. M., et al. (2000). Identification of Novel MAP Kinase Pathway Signaling Targets by Functional Proteomics and Mass Spectrometry. *Molecular Cell*, *6*(6), 1343–1354. doi:10.1016/S1097-2765(00)00132-5

145. Germann, S., Gratadou, L., Dutertre, M., & Auboeuf, D. (2012). Splicing Programs and Cancer. *Journal of Nucleic Acids*, *2012*, 1–9. doi:10.1155/2012/269570

146. Abril, J. F., Castelo, R., & Guigó, R. (2005). Comparison of splice sites in mammals and chicken. *Genome research*, *15*(1), 111–119. doi:10.1101/gr.3108805

147. Turunen, J. J., Niemelä, E. H., Verma, B., & Frilander, M. J. (2012). The significant other: splicing by the minor spliceosome. *Wiley interdisciplinary reviews. RNA*, *4*(1), 61–76. doi:10.1002/wrna.1141

148. Thomassen, M., Blanco, A., Montagna, M., Hansen, T. V. O., Pedersen, I. S., Gutiérrez-Enríquez, S., et al. (2012). Characterization of BRCA1 and BRCA2 splicing variants: a collaborative report by ENIGMA consortium members. *Breast cancer research and treatment*, *132*(3), 1009–1023. doi:10.1007/s10549-011-1674-0

149. Rogan, P. K., Faux, B. M., & Schneider, T. D. (1998). Information analysis of human splice site mutations. *Human Mutation*, *12*(3), 153–171. doi:10.1002/(SICI)1098-1004(1998)12:3<153::AID-HUMU3>3.0.CO;2-I

150. Nalla, V. K., & Rogan, P. K. (2005). Automated splicing mutation analysis by information theory. *Human Mutation*, *25*(4), 334–342. doi:10.1002/humu.20151

151. Mucaki, E. J., Shirley, B. C., & Rogan, P. K. (2013). Prediction of Mutant mRNA Splice Isoforms by Information Theory-Based Exon Definition. *Human Mutation*, n/a–n/a. doi:10.1002/humu.22277

152. Shirley, B. C., Mucaki, E. J., Whitehead, T., Costea, P. I., Akan, P., & Rogan, P. K. (2013). Interpretation, Stratification and Evidence for Sequence Variants Affecting mRNA Splicing in Complete Human Genome Sequences. *Genomics, Proteomics & Bioinformatics*, *11*(2), 77–85. doi:10.1016/j.gpb.2013.01.008

153. Mohapatra, B., Casey, B., Li, H., Ho-Dawson, T., Smith, L., Fernbach, S. D., et al. (2008). Identification and functional characterization of NODAL rare variants in heterotaxy and isolated cardiovascular malformations. *Human Molecular Genetics*, 1–11. doi:10.1093/hmg/ddn411

154. Keren, H., Lev-Maor, G., & Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics*, *11*(5), 345–355. doi:10.1038/nrg2776

155. Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., et al. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA*, *19*(2), 141–157. doi:10.1261/rna.035667.112

156. Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. doi:10.1038/nature07509

157. Shapiro, I. M., Cheng, A. W., Flytzanis, N. C., Balsamo, M., Condeelis, J. S., Oktay, M. H., et al. (2011). An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS Genetics*, *7*(8), e1002218. doi:10.1371/journal.pgen.1002218

158. Thiery, J. P., Acloque, H., Huang, R. Y. J., & Nieto, M. A. (2009). Epithelial-mesenchymal transitions in development and disease. *Cell*, *139*(5), 871–890. doi:10.1016/j.cell.2009.11.007

159. Harper, S. J., & Bates, D. O. (2008). VEGF-A splicing: the key to anti-angiogenic therapeutics? *Nature Reviews Cancer*, *8*(11), 880–887. doi:10.1038/nrc2505

160. Ladomery, M. R., Harper, S. J., & Bates, D. O. (2007). Alternative splicing in angiogenesis: The vascular endothelial growth factor paradigm. *Cancer Letters*, *249*(2), 133–142. doi:10.1016/j.canlet.2006.08.015

161. Nowak, D. G., Woolard, J., & Amin, E. M. (2008). Expression of pro-and anti-angiogenic isoforms of VEGF is differentially regulated by splicing and growth factors. *Journal of cell …*.

162. Pritchard-Jones, R. O., Dunn, D. B. A., Qiu, Y., Varey, A. H. R., Orlando, A., Rigby, H., et al. (2007). Expression of VEGFxxxb, the inhibitory isoforms of VEGF, in malignant melanoma. *British Journal of Cancer*, *97*(2), 223–230. doi:10.1038/sj.bjc.6603839

163. Yeo, G. W., Xu, X., Liang, T. Y., Muotri, A. R., Carson, C. T., Coufal, N. G., & Gage, F. H. (2007). Alternative Splicing Events Identified in Human Embryonic Stem Cells and Neural Progenitors. *PLoS computational biology*, *3*(10), e196–17. doi:10.1371/journal.pcbi.0030196

164. Salomonis, N., Nelson, B., Vranizan, K., Pico, A. R., Hanspers, K., Kuchinsky, A., et al. (2009). Alternative Splicing in the Differentiation of Human Embryonic Stem Cells into Cardiac Precursors. *PLoS computational biology*, *5*(11), e1000553–17. doi:10.1371/journal.pcbi.1000553

165. Gabut, M., Samavarchi-Tehrani, P., Wang, X., Slobodeniuc, V., O'Hanlon, D., Sung, H.-K., et al. (2011). An Alternative Splicing Switch Regulates Embryonic Stem Cell Pluripotency and Reprogramming. *Cell*, *147*(1), 132–146. doi:10.1016/j.cell.2011.08.023

166. Takeda, J., Seino, S., & Bell, G. I. (1992). Human Oct3 gene family: cDNA sequences, alternative splicing, gene organization, chromosomal location, and expression at low levels in adult tissues. *Nucleic Acids Research*, *20*(17), 4613–4620. doi:10.1093/nar/20.17.4613

167. Atlasi, Y., Mowla, S. J., Ziaee, S. A. M., Gokhale, P. J., & Andrews, P. W. (2008). OCT4 Spliced Variants Are Differentially Expressed in Human Pluripotent and Nonpluripotent Cells. *STEM CELLS*, *26*(12), 3068–3074. doi:10.1634/stemcells.2008-0530

168. Das, S., Jena, S., & Levasseur, D. N. (2011). Alternative splicing produces Nanog protein variants with different capacities for self-renewal and pluripotency in embryonic stem cells. *Journal of Biological Chemistry*, *286*(49), 42690–42703.

169. Kim, J. S., Kim, J., Kim, B. S., Chung, H. Y., Lee, Y. Y., Park, C. S., et al. (2005). Identification and functional characterization of an alternative splice variant within the fourth exon of human nanog. *Experimental & molecular medicine*, *37*(6), 601–607. doi:10.1038/emm.2005.73

170. Fackenthal, J. D., & Godley, L. A. (2008). Aberrant RNA splicing and its functional consequences in cancer cells. *Disease Models and Mechanisms*, *1*(1), 37–42. doi:10.1242/dmm.000331

171. Srebrow, A., & Kornblihtt, A. R. (2006). The connection between splicing and cancer. *Journal of Cell Science*, *119*(Pt 13), 2635–2641. doi:10.1242/jcs.03053

172. Holm, F., Hellqvist, E., Mason, C. N., Ali, S. A., Delos-Santos, N., Barrett, C. L., et al. (2015). Reversion to an embryonic alternative splicing program enhances leukemia stem cell self-renewal. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(50), 15444–15449. doi:10.1073/pnas.1506943112

173. Dales, J.-P., Beaufils, N., Silvy, M., Picard, C., Pauly, V., Pradel, V., et al. (2010). Hypoxia inducible factor 1alpha gene (HIF-1alpha) splice variants: potential prognostic biomarkers in breast cancer. *BMC medicine*, *8*, 44. doi:10.1186/1741-7015-8-44

174. Huang, C.-S., Shen, C.-Y., Wang, H.-W., Wu, P.-E., & Cheng, C.-W. (2007). Increased expression of SRp40 affecting CD44 splicing is associated with the clinical outcome of lymph node metastasis in human breast cancer. *Clinica chimica acta; international journal of clinical chemistry*, *384*(1-2), 69–74. doi:10.1016/j.cca.2007.06.001

175. Tress, M. L., Abascal, F., & Valencia, A. (2016). Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends in Biochemical Sciences*, 1–13. doi:10.1016/j.tibs.2016.08.008

176. Hegyi, H., Kalmar, L., Horvath, T., & Tompa, P. (2011). Verification of alternative splicing variants based on domain integrity, truncation length and intrinsic protein disorder. *Nucleic Acids Research*, *39*(4), 1208–1219. doi:10.1093/nar/gkq843

177. Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., et al. (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*, *164*(4), 805–817. doi:10.1016/j.cell.2016.01.029

178. Buljan, M., Chalancon, G., Eustermann, S., Wagner, G. P., Fuxreiter, M., Bateman, A., & Babu, M. M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks.

Molecular Cell, *46*(6), 871–883. doi:10.1016/j.molcel.2012.05.039

179. Ellis, J. D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J. A., et al. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular Cell*, *46*(6), 884–892. doi:10.1016/j.molcel.2012.05.037

180. Haynes, C., Oldfield, C. J., Ji, F., Klitgord, N., Cusick, M. E., Radivojac, P., et al. (2006). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS computational biology*, *2*(8), e100. doi:10.1371/journal.pcbi.0020100

181. Merkin, J., Russell, C., Chen, P., & Burge, C. B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, *338*(6114), 1593–1599. doi:10.1126/science.1228186

182. Hoeben, A. (2004). Vascular Endothelial Growth Factor and Angiogenesis. *Pharmacological Reviews*, *56*(4), 549–580. doi:10.1124/pr.56.4.3

183. Bevan, H. S., van den Akker, N. M. S., Qiu, Y., Polman, J. A. E., Foster, R. R., Yem, J., et al. (2008). The Alternatively Spliced Anti-Angiogenic Family of VEGF Isoforms $VEGF_{xxx}b$ in Human Kidney Development. *Nephron Physiology*, *110*(4), p57–p67. doi:10.1159/000177614

184. Woolard, J. (2004). VEGF165b, an Inhibitory Vascular Endothelial Growth Factor Splice Variant: Mechanism of Action, In vivo Effect On Angiogenesis and Endogenous Protein Expression. *Cancer Research*, *64*(21), 7822–7835. doi:10.1158/0008-5472.CAN-04-0934

185. Elkon, R., Ugalde, A. P., & Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. *Nature Reviews Genetics*, *14*(7), 496–506. doi:10.1038/nrg3482

186. Carswell, S., & Alwine, J. C. (1989). Efficiency of utilization of the simian virus 40 late polyadenylation site: effects of upstream sequences. *Molecular and Cellular Biology*.

187. Wickens, M., & Stephenson, P. (1984). Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3'end formation. *Science*.

188. Gil, A., & Proudfoot, N. J. (1984). A sequence downstream of AAUAAA is required for rabbit |[beta]|-globin mRNA 3|[prime]|-end formation. *Nature*, *312*(5993), 473–474. doi:10.1038/312473a0

189. Gil, A., & Proudfoot, N. J. (1987). Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit β-globin mRNA 3′ end

formation. *Cell*, *49*(3), 399–406. doi:10.1016/0092-8674(87)90292-3

190. Retelska, D., Iseli, C., Bucher, P., Jongeneel, C. V., & Naef, F. (2006). BMC Genomics. *BMC Genomics*, *7*(1), 176–10. doi:10.1186/1471-2164-7-176

191. Zhang, X. H. F., Leslie, C. S., & Chasin, L. A. (2005). Computational searches for splicing signals. *Methods*, *37*(4), 292–305. doi:10.1016/j.ymeth.2005.07.011

192. Ji, Z., Lee, J. Y., Pan, Z., Jiang, B., & Tian, B. (2009). Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(17), 7028–7033. doi:10.1073/pnas.0900028106

193. Mayr, C., & Bartel, D. P. (2009). Widespread Shortening of 3′UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell*, *138*(4), 673–684. doi:10.1016/j.cell.2009.06.016

194. Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., & Burge, C. B. (2008). Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites. *Science*, *320*(5883), 1643–1647. doi:10.1126/science.1155390

195. Ji, Z., & Tian, B. (2009). Reprogramming of 3′ Untranslated Regions of mRNAs by Alternative Polyadenylation in Generation of Pluripotent Stem Cells from Different Cell Types. *PLoS ONE*, *4*(12), e8419–13. doi:10.1371/journal.pone.0008419

196. Wight, M., & Werner, A. (2013). The functions of natural antisense transcripts. *Essays In Biochemistry*, *54*, 91–101. doi:10.1042/bse0540091

197. Khorkova, O., Myers, A. J., Hsiao, J., & Wahlestedt, C. (2014). Natural antisense transcripts. *Human Molecular Genetics*, *23*(R1), R54–R63. doi:10.1093/hmg/ddu207

198. Hunter, T. (2009). Tyrosine phosphorylation: thirty years and counting. *Current Opinion in Cell Biology*, *21*(2), 140–146. doi:10.1016/j.ceb.2009.01.028

199. Deribe, Y. L., Pawson, T., & Dikic, I. (2010). Post-translational modifications in signal integration. *Nature Structural & Molecular Biology*, *17*(6), 666–672. doi:10.1038/nsmb.1842

200. Wang, Y.-C., Peterson, S. E., & Loring, J. F. (2013). Protein post-translational modifications and regulation of pluripotency in human stem cells. *Cell Research*, *24*(2), 143–160. doi:10.1038/cr.2013.151

201. Schwarz, F., & Aebi, M. (2011). Mechanisms and principles of N-linked protein glycosylation. *Current Opinion in Structural Biology*, *21*(5), 576–582.

doi:10.1016/j.sbi.2011.08.005

202. Mitra, N., Sinha, S., Ramya, T. N. C., & Surolia, A. (2006). N-linked oligosaccharides as outfitters for glycoprotein folding, form and function. *Trends in Biochemical Sciences*, *31*(3), 156–163. doi:10.1016/j.tibs.2006.01.003

203. Fusetti, F., Schröter, K. H., Steiner, R. A., van Noort, P. I., Pijning, T., Rozeboom, H. J., et al. (2002). Crystal Structure of the Copper-Containing Quercetin 2,3-Dioxygenase from Aspergillus japonicus. *Structure*, *10*(2), 259–268. doi:10.1016/S0969-2126(02)00704-9

204. Brunner, A. M., Lioubin, M. N., Marquardt, H., Malacko, A. R., Wang, W. C., Shapiro, R. A., et al. (1992). Site-directed mutagenesis of glycosylation sites in the transforming growth factor-beta 1 (TGF beta 1) and TGF beta 2 (414) precursors and of cysteine residues within mature TGF beta 1: effects on secretion and bioactivity. *Molecular endocrinology (Baltimore, Md.)*, *6*(10), 1691–1700. doi:10.1210/mend.6.10.1448117

205. Blanchet, M.-H., Le Good, J. A., Mesnard, D., Oorschot, V., Baflast, S., Minchiotti, G., et al. (2008). Cripto recruits Furin and PACE4 and controls Nodal trafficking during proteolytic maturation. *The EMBO Journal*, *27*(19), 2580–2591. doi:10.1038/emboj.2008.174

206. Le Good, J. A., Joubin, K., Giraldez, A. J., Ben-Haim, N., Beck, S., Chen, Y., et al. (2005). Nodal Stability Determines Signaling Range. *Current Biology*, *15*(1), 31–36. doi:10.1016/j.cub.2004.12.062

207. Donnelly, P., Gabriel, S. B., Green, E. D., Hurles, M. E., Knoppers, B. M., Marth, G. T., et al. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. doi:10.1038/nature15393

208. Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, *9*(6), 477–485. doi:10.1038/nrg2361

209. Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(23), 9362–9367. doi:10.1073/pnas.0903103106

210. Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., et al. (2014). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*. doi:10.1038/nature13835

211. Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., et al. (2014). Defining functional DNA elements in the human genome.

Proceedings of the National Academy of Sciences, *111*(17), 6131–6138. doi:10.1073/pnas.1318948111

212. ENCODE Project Consortium. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology*, *9*(4), e1001046. doi:10.1371/journal.pbio.1001046

213. Majewski, J., & Pastinen, T. (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends in Genetics*, *27*(2), 72–79. doi:10.1016/j.tig.2010.10.006

214. Gebbia, M., Ferrero, G. B., Pilia, G., Bassi, M. T., Aylsworth, A., Penman-Splitt, M., et al. (1997). X-linked situs abnormalities result from mutations in ZIC3. *Nature genetics*, *17*(3), 305–308. doi:10.1038/ng1197-305

215. Roessler, E., Pei, W., Ouspenskaia, M. V., Karkera, J. D., Veléz, J. I., Banerjee-Basu, S., et al. (2009). Cumulative ligand activity of NODAL mutations and modifiers are linked to human heart defects and holoprosencephaly. *Molecular Genetics and Metabolism*, *98*(1-2), 225–234. doi:10.1016/j.ymgme.2009.05.005

216. Bamford, R. N., Roessler, E., Burdine, R. D., Saplakoğlu, U., Cruz, dela, J., Splitt, M., et al. (2000). Loss-of-function mutations in the EGF-CFC gene CFC1 are associated with human left-right laterality defects. *Nature genetics*, *26*(3), 365–369. doi:10.1038/81695

217. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, *337*(6096), 816–821. doi:10.1126/science.1225829

218. Christian, M., Cermak, T., Doyle, E. L., Schmidt, C., Zhang, F., Hummel, A., et al. (2010). Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics*, *186*(2), 757–761. doi:10.1534/genetics.110.120717

219. Joung, J. K., & Sander, J. D. (2012). TALENs: a widely applicable technology for targeted genome editing. *Nature reviews. Molecular cell biology*, *14*(1), 49–55. doi:10.1038/nrm3486

220. Sun, N., & Zhao, H. (2013). Transcription activator-like effector nucleases (TALENs): A highly efficient and versatile tool for genome editing. *Biotechnology and Bioengineering*, n/a–n/a. doi:10.1002/bit.24890

221. Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, *8*(11), 2281–2308. doi:10.1038/nprot.2013.143

222. Doudna, J. A., & Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science*, *346*(6213), 1258096–1258096.

doi:10.1126/science.1258096

223. Ochiai, H., Miyamoto, T., Kanai, A., Hosoba, K., Sakuma, T., Kudo, Y., et al. (2014). TALEN-mediated single-base-pair editing identification of an intergenic mutation upstream of BUB1B as causative of PCS (MVA) syndrome. *Proceedings of the National Academy of Sciences*, *111*(4), 1461–1466. doi:10.1073/pnas.1317008111

224. Spisák, S., Lawrenson, K., Fu, Y., Csabai, I., Cottman, R. T., Seo, J.-H., et al. (2015). CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. *Nature Medicine*, *21*(11), 1357–1363. doi:10.1038/nm.3975

225. Kim, Y., Kweon, J., & Kim, J.-S. (2013). TALENs and ZFNs are associated with different mutation signatures. *Nature Methods*, *10*(3), 185. doi:10.1038/nmeth.2364

226. Hockemeyer, D., & Jaenisch, R. (2016). Induced Pluripotent Stem Cells Meet Genome Editing. *Cell Stem Cell*, *18*(5), 573–586. doi:10.1016/j.stem.2016.04.013

227. Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. doi:10.1038/nature07509

Chapter 2

2 Characterization of a functional non-coding *NODAL* single nucleotide polymorphism (SNP)

2.1 Introduction

2.1.1 Genetics of human pluripotent stem cells

In recent years, a growing body of literature has focused on genomic instability and the accumulation of copy number alterations that occur within human embryonic stem cell (hESC) lines [1-3]. However, no work has addressed how inherited genetic variation is associated with hESC pluripotency, or any other characteristics of this cell type. Such findings will be crucial to achieving the International Stem Cell Initiative's (ISCI) goal of understanding heterogeneity in human embryonic stem cell line models to potentiate generalizable discoveries [4-6]. It has been suggested that modeling pluripotency with cell lines of diverse genetic ancestries will be necessary to achieve this goal [7, 8]. Despite this realization, the two most commonly studied hESC lines (H9 and H1) appear in more publications than the next 20 most common hESC lines combined, and account for over 25% of all hESC citations (http://www.umassmed.edu/iscr/). Thus, genetic polymorphisms in these and other cell lines likely contribute to bias in our current understanding of human pluripotency and early embryonic development.

The genome-wide impact of genetic heterogeneity on gene expression for established hES cell lines is confounded by differences in their derivation. However, this impact has been examined in induced pluripotent stem (iPS) cells where derivation of multiple lines in parallel can be carefully controlled [9, 10]. Strikingly, germ-line genetic variation between individual donors was found to explain more variance in gene expression than the somatic cell type used for reprogramming. Genetic variation has also been implicated in subsequent differentiation potential of iPS cells [11, 12]. Still, beyond the general impact of genetic variation on gene expression profiles, no inherited polymorphisms have been associated with any characteristics of human pluripotent stem cells.

2.1.2 *NODAL* in human pluripotent stem cells

One gene that plays an important role in determining hES cell fate is the TGF-beta superfamily member nodal growth differentiation factor (*NODAL*). In hESCs, NODAL signalling helps maintain pluripotency, partially through transcriptional activation of the transcription factor *NANOG* [13]. NODAL also activates gene expression from poised epigenetic marks, facilitating early differentiation events [14, 15]. To date, only two common single nucleotide polymorphisms (SNPs) in the *NODAL* gene have been functionally studied, along with numerous rare disease-associated mutations [16]—all of which are found in protein coding regions of *NODAL*. However, it is currently unknown how any genetic polymorphisms at the *NODAL* gene locus impact hESC biology. Furthermore, no non-coding *NODAL* SNPs have ever been functionally characterized in any context. Here I explore the associations and functional impact of a non-coding intronic *NODAL* SNP (rs2231947) in hES cell lines.

2.2 Results

Using SNP genotyping data from the International Stem Cell Initiative's (ISCI's) global survey of hESC lines [5], and associated gene expression data [4], I discovered two interesting associations in hES cell lines for *NODAL* SNP rs2231947. The relative location of rs2231947 is shown in the context of the human *NODAL* gene in Figure 2.1. First, I found the minor allele for rs2231947 (T on the sense strand) to be drastically under-represented in male hESC lines of European ancestry relative to ancestry-matched female hESC lines (Figure 2.2). The association between rs2231947 genotype and an individual's sex was not present in the European reference super population from the 1000 Genomes Project, suggesting this bias does not occur under normal developmental conditions. Furthermore, the minor allele frequency (MAF) for rs2231947 in *female* hESC lines did not differ from that of the European reference super population, suggesting that prospective male cell lines with the minor allele for rs2231947 may have been negatively selected against. The sex association was not due to an ancestry stratification effect, as analysis of all five available European subpopulations showed extremely low differentiation for rs2231947 (Table 2.1).

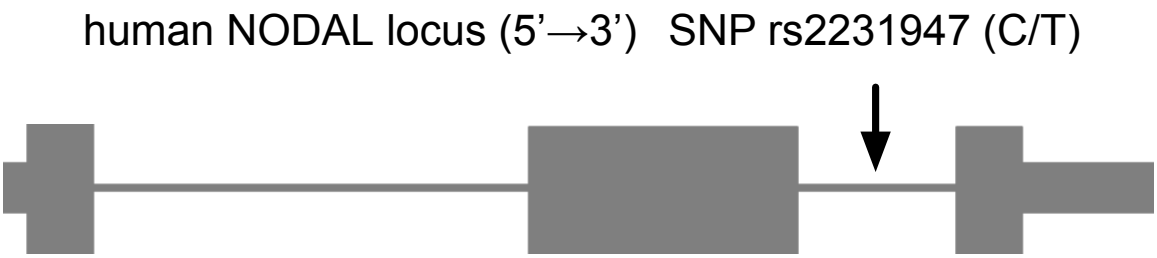


Figure 2.1: **Schematic of the human *NODAL* gene locus on chromosome 10.** Orientation is based on the sense strand, with the 5' end on the left and 3' end on the right. Thick bars indicate coding regions, intermediate bars indicate untranslated regions, and thin lines indicate introns. The approximate position of single nucleotide polymorphism (SNP) rs2231947 is indicated. Diagram scale is approximate.

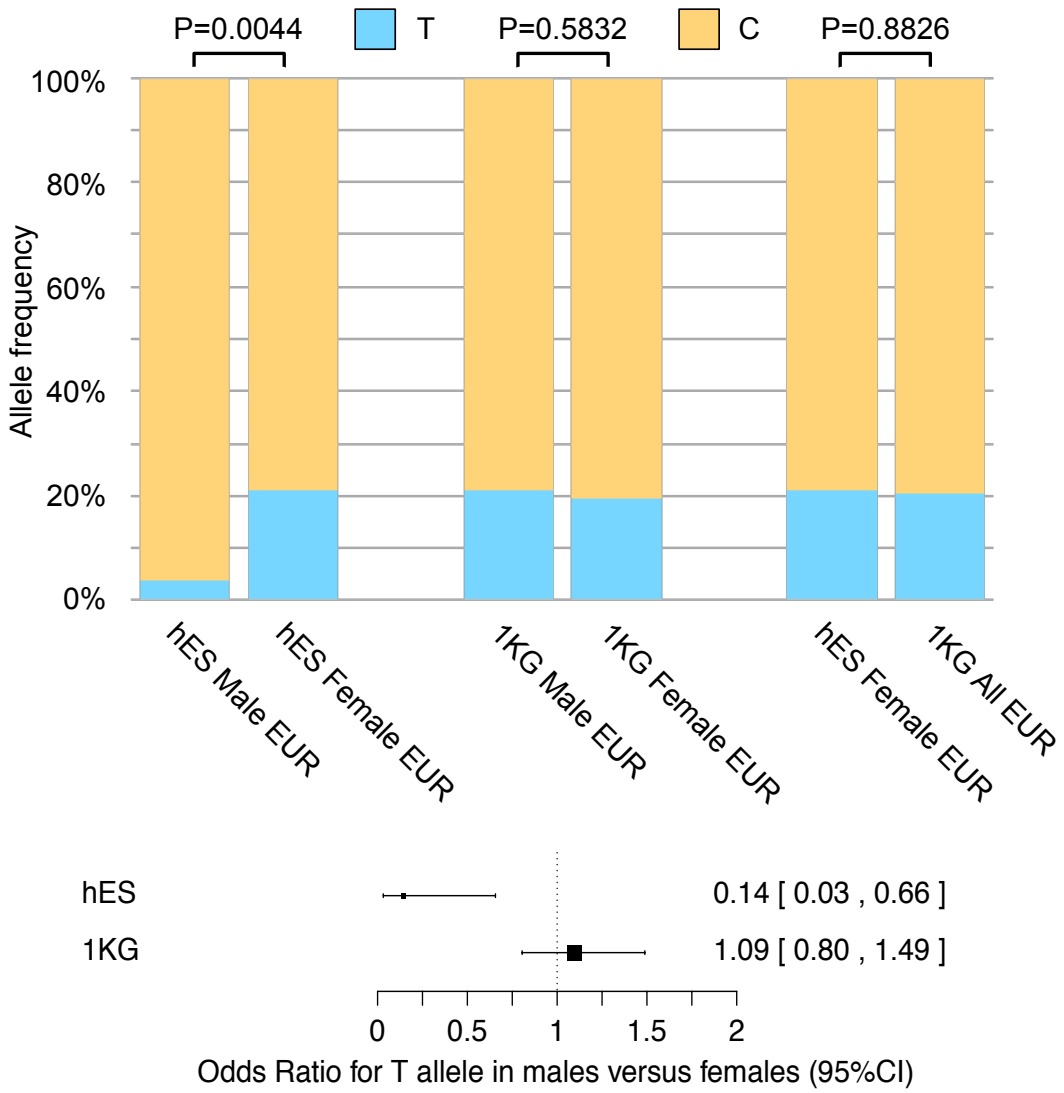


Figure 2.2: *NODAL* SNP rs2231947 sex bias in hES cell lines.
Upper: From left to right: rs2231947 allele frequencies in male (T=2, C=52, n=54) and female (T=16, C=60, n=76) hES cell lines, male (T=101, C=379, n=480) and female (T=103, C=423, n=526) individuals from the 1000 Genomes Project (1KG), and female (n=76) hES cell lines and all (male and female, n=1,006) individuals from the 1000 Genomes Project. n=number of alleles. All cell lines and individuals are of European (EUR) ancestry (see methods). P values for each pair indicate results of two-tailed Fisher exact tests. Bottom: Forest plot of the odds ratio (OR) for a cell line or individual having the T allele for rs2231947 in males versus females. Black square indicates OR, lines indicate 95% confidence interval (CI). Numbers to the right are OR [minimum of CI, maximum of CI].

Table 2.1: Extent of genetic differentiation among European subpopulations for SNP rs2231947.

Populations compared	rs2231947 F_{st}
CEU_FIN	0.004693
CEU_GBR	0.000016
CEU_IBS	-0.004606
CEU_TSI	-0.004883
FIN_GBR	-0.004823
FIN_IBS	0.004043
FIN_TSI	0.003740
GBR_IBS	-0.000386
GBR_TSI	-0.000707
IBS_TSI	-0.004676

Weir and Cockerham's F_{st} was calculated for each pair of European subpopulations from the 1000 Genomes Project. This metric is a measure of the extent to which two populations are genetically different, and generally ranges from 0 (identical allele frequencies) to 1 (complete allele switching). All comparisons shown here are very close to 0, suggesting there is very little differentiation at the rs2231947 locus between European subpopulations. Note that this method for calculating F_{st} may yield slightly negative values, but such values have no biological meaning. Population codes: CEU= Utah Residents (CEPH) with Northern and Western European Ancestry, FIN= Finnish in Finland, GBR= British in England and Scotland, IBS= Iberian Population in Spain, TSI= Toscani in Italia.

Since rs2231947 alleles showed a sex bias in human embryonic stem cells, I hypothesized that rs2231947 genotype may correlate with sex-specific gene expression. I chose to analyze X-inactive specific transcript (XIST), a major driver of the female-specific X-chromosome inactivation (XCI) process that takes place in early embryonic development [17]. In female hESC lines (n=17), I found that the rs2231947 T allele had a strong positive association with XIST expression. Female T|T or C|T (n=5) cell lines expressed XIST transcript at a median level of 1,648-fold higher than the median of C|C (n=12) cell lines (P=0.015, Figure 2.3).

To assess a potential function for SNP rs2231947 and assess its contribution to *NODAL* biology, I first examined the sequence context of the rs2231947 locus. When the T allele is present, the locus closely resembles a typical human splice site motif (Figure 2.4 and [18]). Conducting more detailed "Automated Splice Site And Exon Definition Analyses" [19] revealed that relative to the C allele, the T allele of SNP rs2231947 was predicted to both slightly strengthen a putative splice acceptor site, as well as contribute to a strong cryptic 5' splice donor site not formed by the rs2231947 C allele (Figure 2.4).

Until recently, there was only one annotated *NODAL* transcript isoform (NCBI RefSeq NM_018055.4). During writing, a second isoform was curated into the RefSeq database (NM_001329906.1). These isoforms differ in their use of alternative first exons. No other *NODAL* transcript variants have been described. Based on the bioinformatic splice site predictions, I next conducted RT-PCR to detect any potential novel exons. I designed primers to target constitutive exons 2 and 3 flanking the rs2231947 SNP within intron 2. The H9 hES cell line was chosen for analysis as it was found to be homozygous for the minor T allele of rs2231947 predicted to contribute to a strong cryptic splice donor site. In addition to the expected product corresponding to the primary annotated *NODAL* transcript, a second product was detected. Cloning and sequencing of this amplicon revealed a 116 base-pair cassette exon forming upstream and downstream junctions with the second and third constitutively spliced *NODAL* exons, respectively (Figure 2.5). The 5' splice donor site defining this alternative exon corresponded to the site predicted to be strengthened by the T allele of rs2231947. Next, a panel of hES and human induced

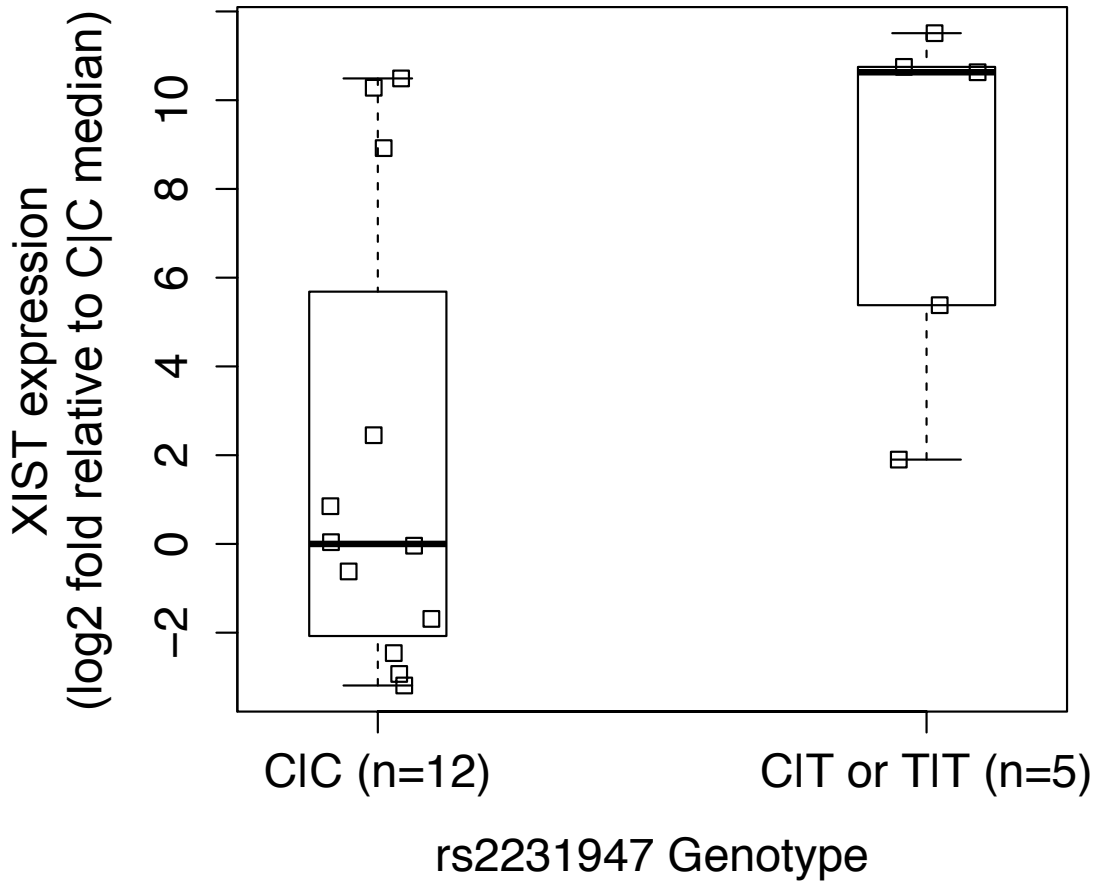


Figure 2.3: *NODAL* SNP rs2231947 genotype is associated with XIST levels in female hES cell lines.

Boxes indicate median and inter-quartile ranges. Whiskers indicate minimum and maximum observations. XIST expression is lower in C|C compared to C|T or T|T female hES cell lines. P=0.015 by one-tailed t test.

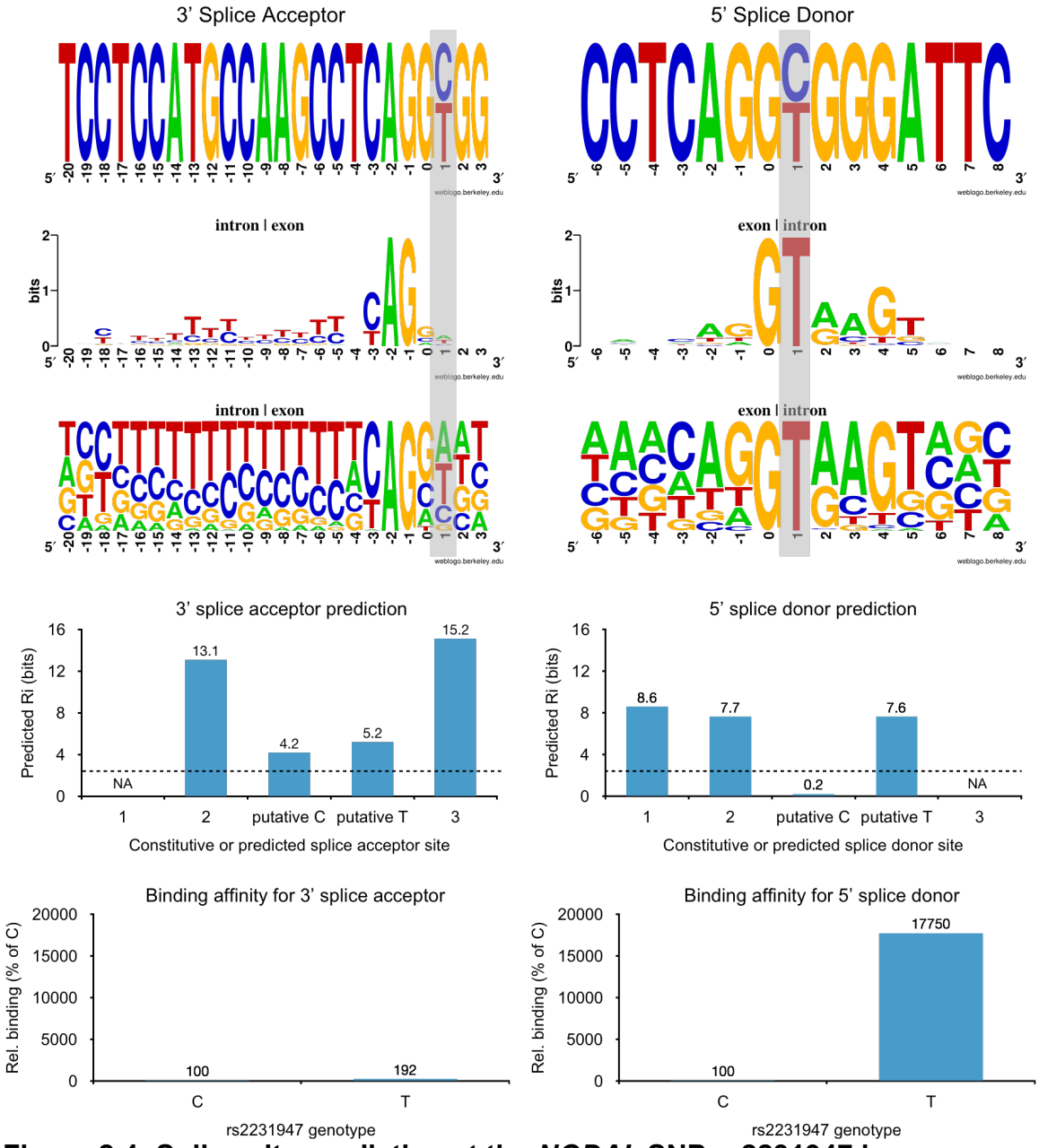


Figure 2.4: Splice site prediction at the *NODAL* SNP rs2231947 locus. Sequence "web logos" show human *NODAL* rs2231947 locus relative to both splice acceptor sites (left) and splice donor sites (right). rs2231947 is shown as C/T SNP in sequences. For splice acceptor sites, position "0" marks the intron-exon boundary and is the first (most 5') base of an exon. For splice donor sites, position "0" marks the exon-intron boundary and is the first (most 5') base of an intron. "Putative C" and "putative T" refer to predicted splice sites at the rs2231947 locus contributed by each SNP allele. Exons 1, 2, and 3 are the constitutively spliced *NODAL* exons. The dashed lines indicate the predicted minimum threshold of 2.4 bits for splice site utilization.

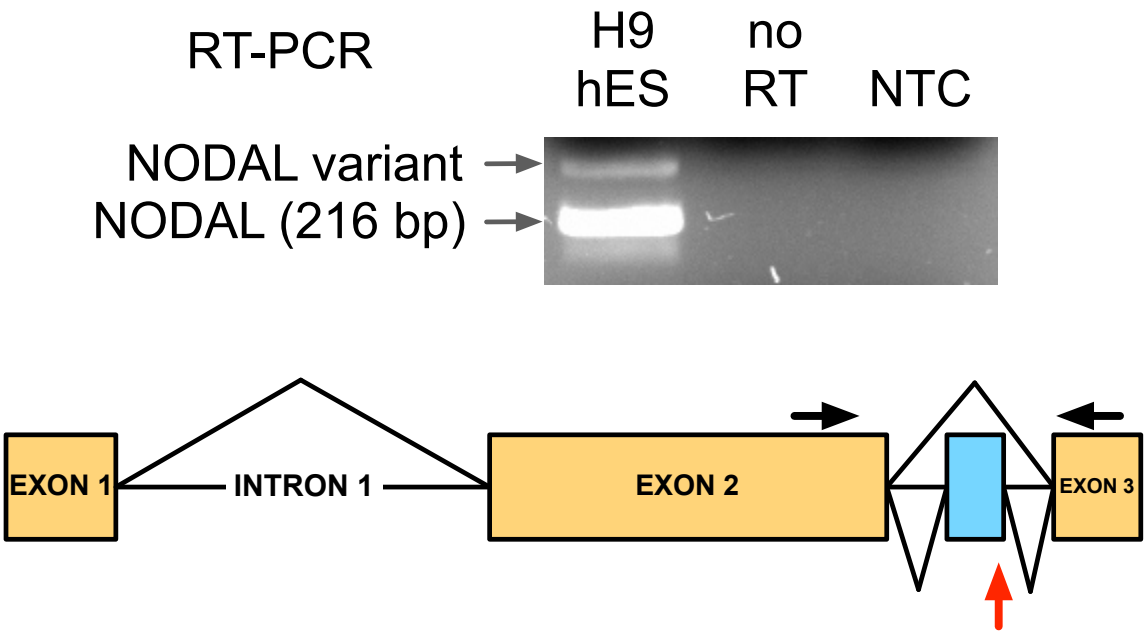


Figure 2.5: Novel *NODAL* transcript isoform in H9 hES cells.
Reverse-transcription PCR amplification of *NODAL* reveals a novel *NODAL* transcript at hg19 coordinates: chr10:72193855-72193971. Constitutively spliced *NODAL* exons are shown in orange. The alternative cassette exon is shown in blue. Black arrows indicate PCR primer sites. Red arrow indicates the SNP rs2231947 locus. Diagonal lines indicate splice site utilization.

pluripotent (iPS) stem cell lines were genotyped for rs2231947 and assayed for both the primary annotated *NODAL* and the new-found *NODAL* variant transcripts using real time PCR assays. While all cell lines expressed the primary annotated *NODAL* transcript, only cell lines with at least one T allele (T|T or C|T genotypes) expressed the novel *NODAL* transcript (Figure 2.6).

Next, a minigene splicing reporter plasmid [20] was modified to include *NODAL* sequence spanning from the 3' region of constitutively spliced exon 2, to the 5' region of constitutively spliced exon 3. Transfection of cells with this minigene plasmid followed by RT-PCR analysis of *NODAL* gene expression specific to the plasmid demonstrated that expression of the two *NODAL* isoforms is characteristic of true alternative splicing from a single locus, as opposed to mutually-exclusive splicing of one isoform or the other based on the SNP allele present in cis. Furthermore, the generation of minigenes with different alleles (C or T) for rs2231947 also revealed that SNP rs2231947 directly regulates this alternative splicing event, and that the SNP rs2231947 T allele is necessary for inclusion of the alternative cassette exon (Figure 2.7).

Allele-specific gene expression can be used to determine to what extent each chromosome/ allele contributes to expression of a given transcript. A significant fraction of genes in human embryonic stem cells display allele-biased gene expression [21]. Since these biases can result from parent-of-origin effects, and I found *NODAL* SNP rs2231947 genotypes displayed a sex bias in hES cells, I was interested in assessing allelic expression of *NODAL* in hES cells. A heterozygous SNP allele in an exon serves as an ideal marker to assess allelic gene expression. A survey for such polymorphisms with relatively high population MAFs (> 5%, and therefore likely to be heterozygous) in *NODAL* exons found three such SNPs, with constitutive exon 2 SNP rs104894169 having the highest MAF. I surveyed a panel of hES cell lines and found the CA1 line to be heterozygous for this SNP. This cell line was also ideal for analysis as it was heterozygous for SNP rs2231947 (Figure 2.6). Sequencing of clones of a PCR amplicon within constitutive exon 2 encompassing the rs104894169 locus amplified from CA1 cDNA revealed expression of both A and G SNP alleles (Figure 2.8A). Similar analysis

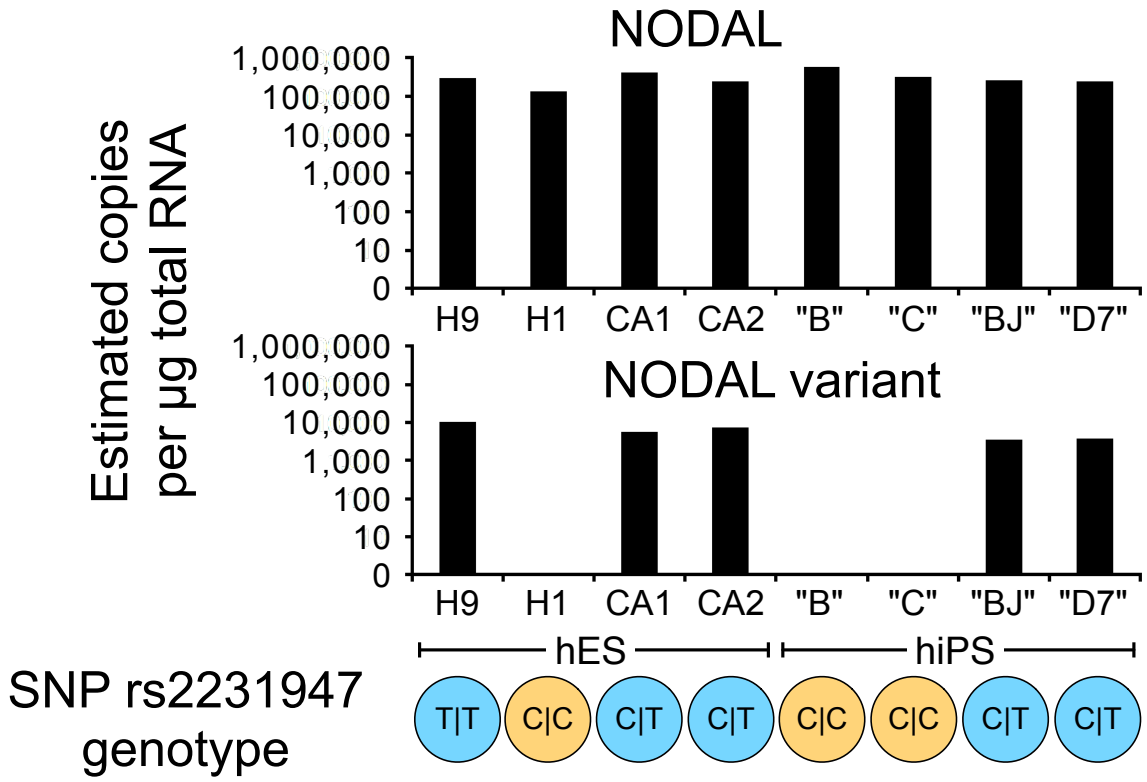


Figure 2.6: *NODAL* variant expression is associated with SNP rs2231947 genotype in human pluripotent stem cell lines.
Real time PCR isoform-specific analysis of *NODAL* reveals *NODAL* variant expression in only a subset of human pluripotent (hES and hiPS) cell lines with at least one T allele at SNP rs2231947 (C|T or T|T, blue). C|C genotypes are indicated in yellow. Cell line codes: “B” = 0901B, “C” = 0901C, “BJ” = BJ10, “D7” = 1681D7. hES= human embryonic stem cell, hiPS= human induced pluripotent stem cell.

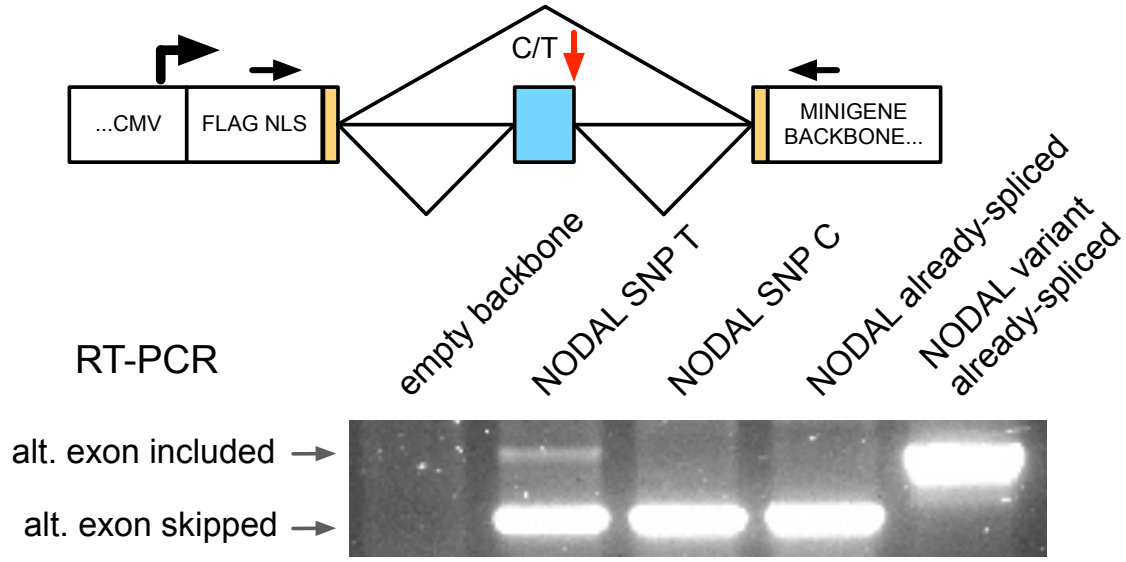


Figure 2.7: The SNP rs2231947 T allele is necessary for alternative splicing of a *NODAL* minigene.

Portions of constitutively spliced *NODAL* exons are shown in orange. The alternative cassette exon is shown in blue. Red arrow indicates rs2231947 position. “Already spliced” refers to plasmids where *NODAL* genomic DNA template has been replaced with the corresponding cDNA amplified using the same primers. Black arrows indicate primer sites used for RT-PCR analysis of minigene splicing. “alt” = alternative.

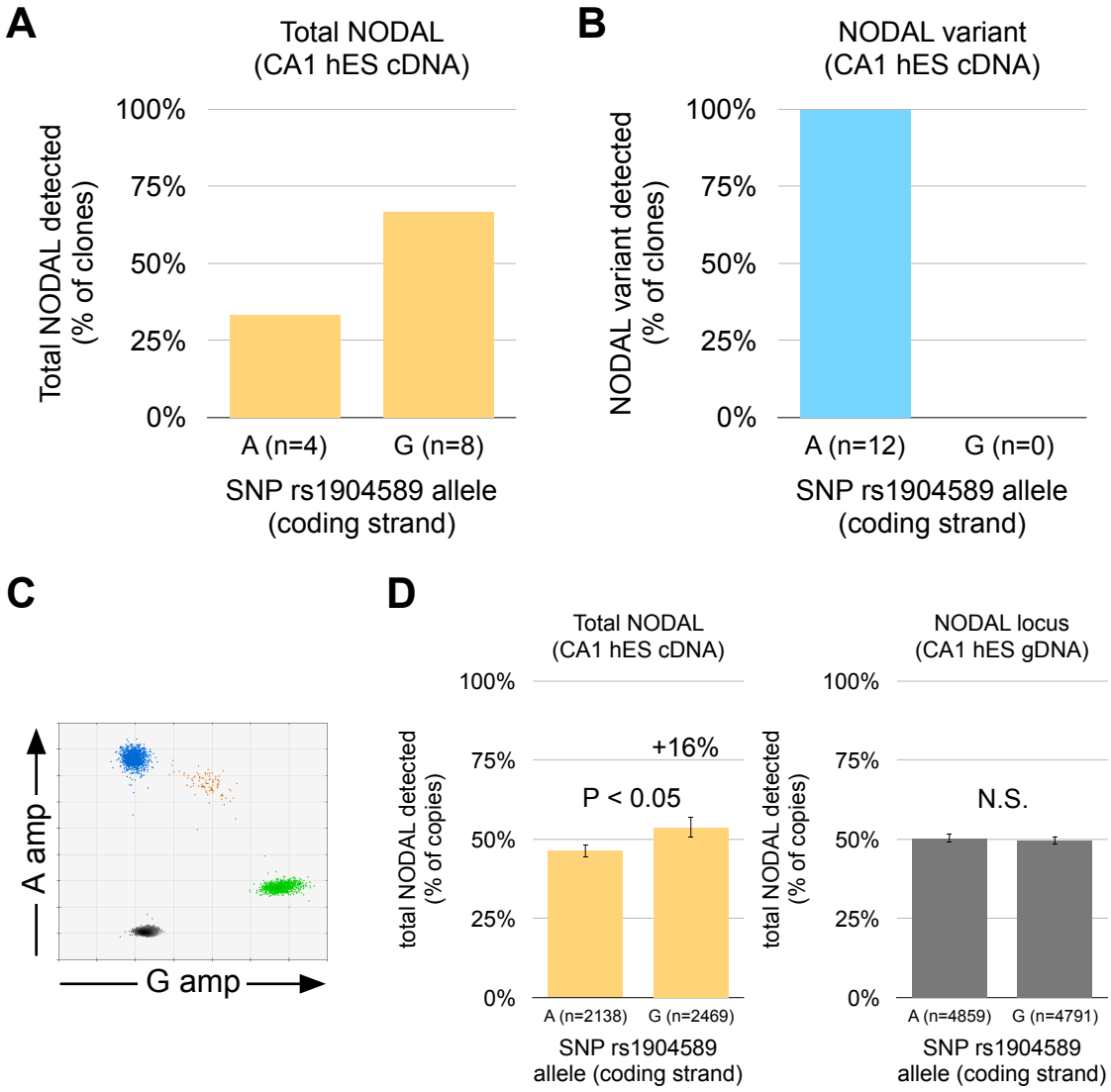


Figure 2.8: *NODAL* expression is biallelic in CA1 hES cells.
A) Both rs19048589 alleles are expressed in total *NODAL* transcript from the heterozygous CA1 hES cell line. B) Only A alleles are expressed in *NODAL* variant transcripts from the CA1 cell line also heterozygous for rs2231947. C) Example of ddPCR results for high throughput detection of allelic expression of *NODAL* transcript. D) Left: quantification of ddPCR results for total *NODAL* transcript. Right: quantification of ddPCR results for genomic DNA copy number baseline.

of rs104894169 alleles in an amplicon specific to *NODAL* variant transcripts revealed expression of only the A allele (Figure 2.8B). This was consistent with heterozygous alleles for rs2239147 and indicated that the A allele for rs104894169 was found on the same chromosome as the T allele for rs2231947. These results also confirmed *endogenous* true alternative splicing of *NODAL*. For total *NODAL* transcript, since more clones were found with the rs104894169 G allele than the A allele, I utilized a more high-throughput approach to determine if *NODAL* expression demonstrated allelic bias in this fashion. Using a droplet digital PCR (ddPCR) SNP genotyping assay for rs104894169 (Figure 2.8C), expression was found to be only slightly biased toward the G allele (16% higher than the A allele, $P<0.05$), and this difference could not be attributed to differences in genomic DNA copy number between the two chromosomes (Figure 2.8D).

The identification of rs2231947 as a functional polymorphism prompted me to fully characterize the genetic variation represented by the *NODAL* splicing SNP rs2231947. I performed linkage disequilibrium (LD) analysis using raw SNP genotyping data for European reference populations from the 1000 Genomes Project [22] using VCFtools [23], followed by functional annotation of obtained variants using the UCSC Genome Browser's Variant Annotation Integrator [24]. Fourteen SNPs were identified as being in high LD ($R^2 > 0.8$) with rs2231947 (Table 2.2 and Figure 2.9). Post-hoc empirical testing of SNPs in lower LD with rs2231947 ($R^2 < 0.8$) genotyped in the hESC line sample [5] did not reveal any statistically significant associations with the sex of hESC lines (Table 2.3). Variant Annotation Integrator results revealed none of the high LD SNPs were within gene coding regions, had ClinVar annotations, or matched NHGRI GWAS hits. However, several of the high LD SNPs were found in well-characterized *NODAL* enhancers upstream of the transcriptional start site (Figure 2.9). These included the asymmetric enhancer (ASE) within a CpG island, the node enhancer (NDE), and the proximal epiblast enhancer (PEE) [described in [25-27]]. ENCODE data also revealed that the PEE contained a transcription factor "hotspot" bound by 18 different transcription factors in H1 hESCs. One SNP in the PEE (rs35210846) was found within 14 of these binding sites, including a NANOG binding site, and was just downstream of a POU5F1 (also known as OCT4) binding site. These two transcription factors are well-documented master-regulators of pluripotency (reviewed in [28]). Therefore, in addition

Table 2.2: SNPs in high LD (R^2 > 0.8) with rs2231947 in each 1000 Genomes Project reference European subpopulation.

CEU	FIN	GBR	IBS	TSI
rs35345134	rs35345134	rs35345134	rs35345134	rs35345134
rs34843983	rs34843983	rs34843983	rs34843983	rs34843983
rs58202646	rs58202646	rs58202646	rs58202646	rs58202646
rs35914122	rs35914122	rs35914122	rs35914122	rs35914122
rs36038032	rs36038032	rs36038032	rs36038032	rs36038032
rs60746183	rs60746183	rs60746183	rs60746183	rs60746183
rs2231947	rs2231947	rs2231947	rs2231947	rs2231947
rs7094345	rs7094345	rs7094345	rs7094345	rs7094345
rs12777854	rs12777854	rs12777854	rs12777854	rs12777854
rs17512976	rs17512976	rs17512976	rs17512976	rs17512976
rs35356045	rs35356045	rs35356045	rs35356045	rs35356045
rs35767814	rs35767814	rs35767814	rs35767814	rs35767814
rs71012206	rs71012206	rs71012206	rs71012206	rs71012206
rs12764201		rs12764201		
rs35210846	rs35210846	rs35210846	rs35210846	rs35210846

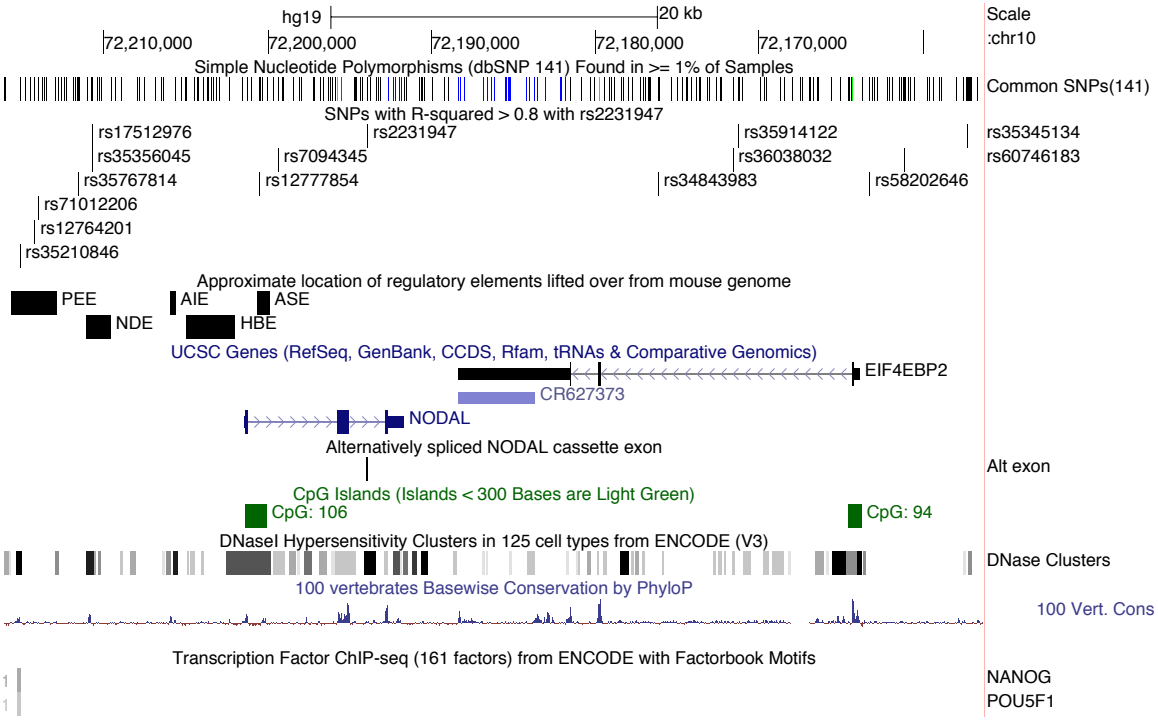


Figure 2.9: SNPs in high LD with rs2231947.

Various tracks from the UCSC genome browser are shown to provide genomic context to the various polymorphisms identified. The image is the reverse of the default orientation for chromosome 10 so that the sense strand of *NODAL* (5'- 3') is shown left to right. Note: only NANOG and POU5F1 (OCT4) hits are shown in the ENCODE transcription factor ChIP-seq track. A "1" beside the ChIP-seq hits indicates presence in the H1 hES cell line. For the *NODAL* regulatory elements, "PEE" = proximal epiblast enhancer, "NDE" = node specific enhancer, "AIE" = asymmetric initiator element, "HBE" = highly bound element, "ASE" = asymmetric enhancer.

Table 2.3: SNPs genotyped in the hES cell line sample with R^2 to rs2231947 < 0.8 do not have alleles represented at different frequencies in male versus female cell lines.

All SNPs with significant LD (R^2 >0.2) to rs2231947 genotyped in the ISCI hES sample are shown. Note that 100% of SNP rs17512976 alleles match rs2231947 alleles (in terms of minor and major allele), as reflected by the high LD between these two SNPs (0.97≤ R^2 ≤1.00).

				R^2 with rs2231947					
SNP	**Minor allele count**	**Major allele count**	**% Match to rs2231947**	**CEU**	**FIN**	**GBR**	**IBS**	**TSI**	**EUR**
rs17512976				0.97	1.00	1.00	1.00	1.00	0.99
Male	2	52	100%						
Female	16	60	100%						
Fisher Test	P = 0.0044								
rs10740348				0.63	0.48	0.64	0.49	0.52	0.55
Male	13	39	58%						
Female	26	46	72%						
Fisher Test	P= 0.2404								
rs7082255				0.57	0.52	0.44	0.57	0.44	0.51
Male	13	41	67%						
Female	23	45	79%						
Fisher Test	P = 0.318								
rs2279253				0.25	0.24	0.22	<0.2	0.21	0.22
Male	24	28	27%						
Female	38	38	55%						
Fisher Test	P = 0.7207								
rs3812706				0.25	0.22	0.23	<0.2	0.20	0.22
Male	22	28	32%						
Female	38	38	55%						
Fisher Test	P = 0.5856								

rs10762381				0.25	0.24	0.23	<0.2	0.22	0.22
Male	24	30	30%						
Female	39	37	55%						
Fisher Test	P = 0.4796								
rs1904589				0.30	0.30	0.26	0.27	0.30	0.29
Male	17	37	52%						
Female	29	43	72%						
Fisher Test	P = 0.353								
rs4607991				0.47	<0.2	0.44	<0.2	0.27	0.26
Male	8	44	69%						
Female	15	61	68%						
Fisher Test	P = 0.6415								
rs10740344				0.32	0.29	0.39	0.27	0.23	0.26
Male	16	38	56%						
Female	27	47	68%						
Fisher Test	P= 0.4534								
rs10740345				0.24	0.22	0.22	<0.2	0.20	0.22
Male	24	28	27%						
Female	37	37	57%						
Fisher Test	P= 0.7194								
rs10823529				0.24	<0.2	0.36	<0.2	<0.2	<0.2
Male	18	34	38%						
Female	30	44	54%						
Fisher Test	P= 0.5776								
rs4570507				0.38	<0.2	0.51	<0.2	0.31	0.26
Male	11	41	62%						
Female	27	47	62%						

Fisher Test	P= 0.0775								
rs7084009				0.30	<0.2	0.36	<0.2	<0.2	<0.2
Male	10	42	65%						
Female	15	61	68%						
Fisher Test	P= 1.0000								

to rs2231947, other polymorphisms in high LD such as rs35210846 may also be functional in hESCs in their effect on *NODAL* gene expression.

I was also interested in determining the degree of conservation of the rs2231947 locus, the splice donor site that encompasses it, and the alternatively spliced cassette exon of *NODAL*. PhyloP conservation scores were obtained for each individual base of interest. Positive numbers correspond to conservation or slower evolution than expected under neutral drift, while negative numbers correspond to accelerated or more rapid evolution than expected under neutral drift [29, 30]. The splice donor site defining the cassette alternative exon shows no highly conserved bases, with the highest scoring base from positions -3 to +5 having a score of only 0.04 (Figure 2.10). For comparison, the splice donor site for *NODAL*'s second exon contains six bases with PhyloP scores greater than 2. Interestingly, the base immediately 3' (on the sense strand) to SNP rs2231947 is very poorly conserved, with a score of -4.94. Of the species with sequence alignments to the rs2231947 locus (all mammals), only one had a base other than C or T (the alleles of human SNP rs2231947). Of the remaining aligned genomes, 5 had a T, and 35 had a C at this position (Figure 2.10). PhyloP scores were also used to profile the typical conservation of a base in the cassette alternative exon relative to the flanking intronic regions (Figure 2.11). Bases in the alternative exon had a significantly higher ($P<0.01$) mean PhyloP score (0.28) than the remainder of the intron which has a score very close to zero (-0.03). In the alternative exon, 64% of bases had a positive score, while in the remainder of the intron that figure was 52%. Notably, the alternatively spliced exon also had a significantly lower average score (0.28) than the highly conserved constitutively spliced second exon of *NODAL* (1.09). Notably, the poorly conserved position immediately adjacent to the rs2231947 locus (Hg19 chr10:72193853) and within the alternative exon splice donor site is predicted to be the most rapidly evolving in the entire intron (Figure 2.11).

Another model for assessing genomic conservation is PhastCons. Instead of the single base independence of PhyloP, this model incorporates the effects of neighbouring bases to identify short runs of conservation or conserved genomic elements [31]. PhastCons

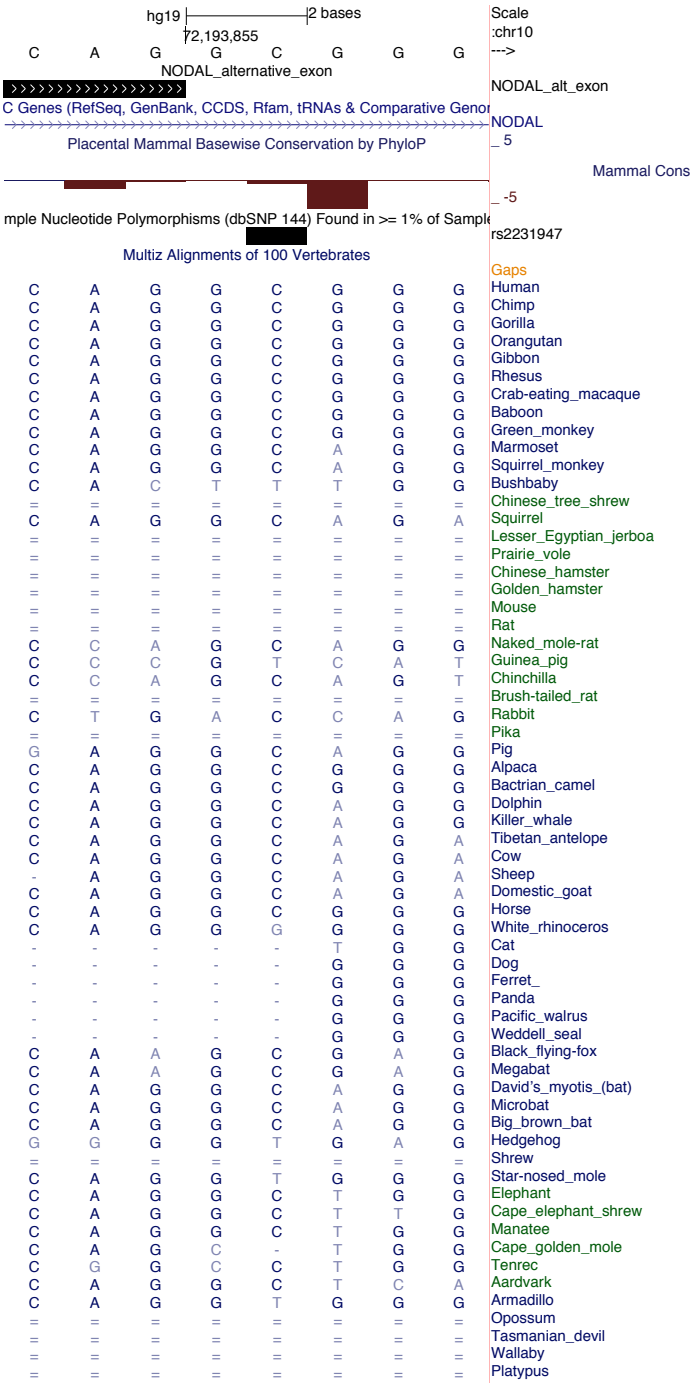


Figure 2.10: Base-wise conservation at the *NODAL* alternative exon splice donor site.

Position -3 to position +5 on the sense strand is shown left to right. PhyloP scores of individual bases are represented graphically. Aligned sequences are shown for mammals. "-" indicates no bases in the aligned region (insertion or deletion). "-" indicates bases that could not be aligned.

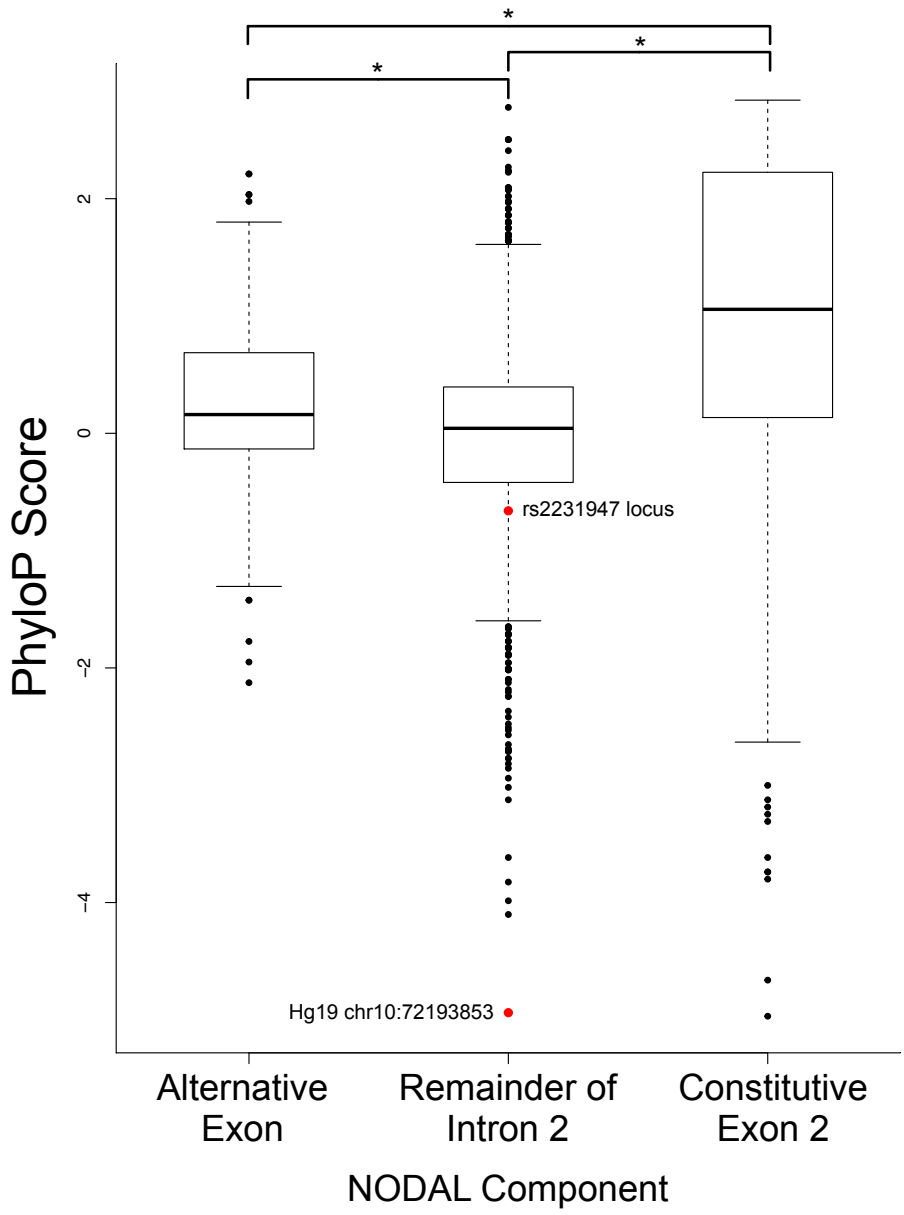


Figure 2.11: PhyloP conservation scores for bases within various *NODAL* elements.

Box edges indicate interquartile ranges. Thick line indicates median values. Dashed lines extend 1.5 interquartile ranges from box, or to maximum value as for constitutive exon 2. Only outliers are shown as individual data points. Red data points highlight rs2231947 position and the adjacent base at hg19 chr10:72193853. * indicates P<0.01 using one-way ANOVA and Tukey HSD Test.

revealed a cluster of three large conservation peaks that are closely bound by the ends of the alternative exon (Figure 2.12).

Population differentiation analysis was conducted to assess differences in allele frequencies for rs2231947 and its linkage group between human populations of different ancestry. High population differentiation was apparent between the East Asian super population and every other super population (African, European, South Asian, and Ad Mixed American). This differentiation was the highest between the East Asian and European super populations (Figure 2.13). The MAF is approximately 1% in the East Asian super population and 20% in the European Super population. Analysis of all similar SNPs on chromosome 10 reveals that rs2231947 is highly differentiated, ranking in the 78th percentile of all similar SNPs (Figure 2.13).

2.3 Discussion

I have shown that the intronic *NODAL* SNP rs2231947 is associated with both XIST expression and the sex of human embryonic stem cell lines. Furthermore, I demonstrated that this SNP is highly functional in modulating the novel alternative splicing of human *NODAL*, resulting in expression of a *NODAL* variant transcript also described here.

The virtual absence of the rs2231947 T allele in male hES cell lines suggests that prospective cell lines of this genetic background were negatively selected, either naturally, or as a consequence of their undesirability for continued use. This selection could have taken place at various stages during cell line derivation, expansion, or continued propagation. It has been previously reported that the derivation of hES cell lines from embryos is a very inefficient process, with only a small fraction of initial embryos used successfully deriving cell lines [32]. Given this inter-embryo variability in the ability to derive established cell lines, the possibility of a genetic influence is unsurprising. Such selection could take place at numerous stages of the process. For example, some prospective cell lines could fail to survive the initial shock of relatively harsh cell culture conditions. Alternatively, some prospective cell lines could be relatively unstable in the pluripotent state and display an undesirable propensity for

Figure 2.12: *NODAL* intron 2 conservation in mammals.
Scale is shown at top of figure. Blue box indicates footprint of *NODAL* alternative exon. For the "multiz alignments" track, vertical lines indicate aligned bases. Horizontal lines indicate gaps in alignments.

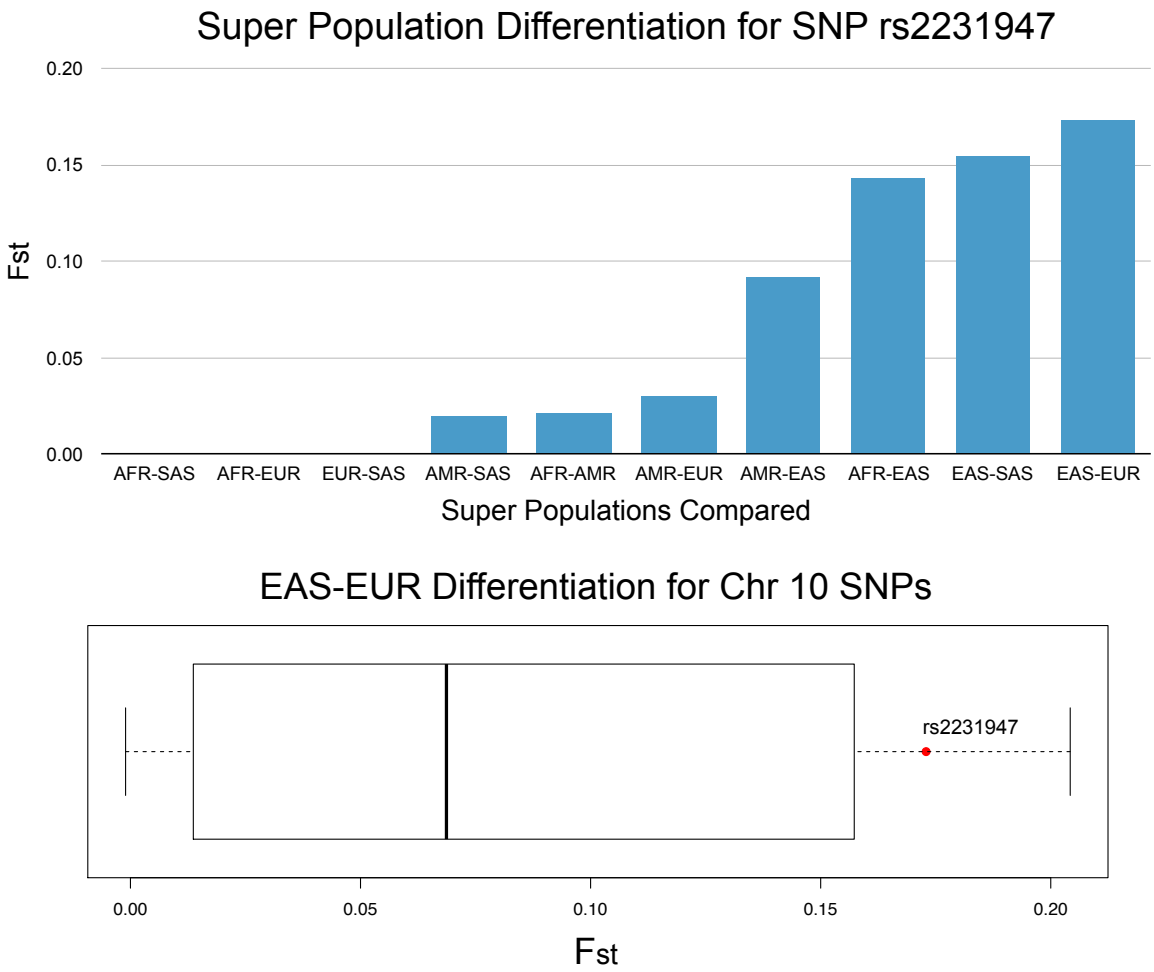


Figure 2.13: Population differentiation analysis for SNP rs2231947.
Top panel: F_{st} is Weir and Cockerham's population differentiation statistic. 1000 Genomes Super Population Codes: AFR= African, EUR= European, SAS= South Asian, AMR= Ad Mixed American, EAS=East Asian. Bottom panel: Box plot of population differentiation between East Asian and European super populations for chromosome 10 SNPs with both European MAF≅0.2 and MAF <0.2 in EAS (n=2,455). Boxes indicate interquartile range. Thick vertical bar indicates median. Dashed lines extend to minimum and maximum values in sample. SNP rs2231947 is noted in red and is in the 78th percentile.

spontaneous differentiation. Such cell lines would have been less likely to be propagated and shared by researchers.

Strikingly, the under-representation of a genetically defined subset of male hESC lines may help explain the previously reported yet still unexplained female bias in established hESC lines [33]. Indeed, the 59% female-to 41% male bias in the ISCI European sample is very similar in magnitude to the female bias reported by Ben-Yosef and colleagues [33]. The genetic association presented here is very strong (odds ratio = 0.14, Figure 2.2). To my knowledge, this is the first genetic association of any kind reported for human embryonic stem cell lines. This is likely due to the single SNP-of-interest approach utilized here. In contrast, a typical genome-wide survey would not be appropriately statistically powered to detect any genetic associations given the number of hES cell lines available. Importantly, the sex association is not due to an ancestry stratification effect, as analysis of all five European reference subpopulations from the 1000 Genomes Project show extremely low differentiation (differences in allele frequencies) for rs2231947 using Weir and Cockerham's F_{st} statistic (Table 2.1). SNP rs2231947 genotyping error is also not a source of the observed bias, as the highly linked SNP rs17512976 shares 100% of genotypes with rs2231947 in the ISCI dataset (Table 2.3). It appears that the sex bias is a cell culture-specific effect and does not appear under normal developmental conditions (Figure 2.2 and [33]). This is unsurprising given that in the embryo, altered or heterogeneous morphogen signals are generally balanced through intricate morphogen gradients, intact feedback loops, and compensating signals [34, 35]. Of course, such compensatory mechanisms are likely completely lost in more homogenous cell culture conditions where growth factors are largely supplied exogenously by culture medium at high doses. Thus, only those cells suited for specific culture conditions are able to be maintained as cell lines in vitro.

In addition to the sex bias, I also found a strong positive association between the rs2231947 T allele and levels of the female specific XIST transcript. This was perhaps surprising given the well-documented heterogeneity and plasticity in both XIST expression and XCI in female hES cell lines, not only between cell lines of different origin, but also between isolates of a single cell line [4, 36, 37]. This result suggests that

genetic variation at the *NODAL* gene locus may explain some of the considerable inter-cell line variability in XIST transcript levels. Indeed, the International Stem Cell Initiative found the inter-cell line variability in XIST levels intriguing as it was highlighted in their profiling of hESC lines [4]. Possible causes of this variability were not offered and have since remained elusive.

Further characterization of rs2231947 revealed that this SNP was functional and had a drastic impact on gene expression from the *NODAL* locus. *In silico* analysis suggested that the SNP possibly modulates overlapping putative splice acceptor and splice donor sites. This was unsurprising given that both splice acceptor and splice donor sites share similar sequence motifs, although with different base-wise conservation profiles (Figure 2.4 and [18]). The T allele for rs2231947 was predicted to affect relative binding affinity at the splice donor site more strongly (Figure 2.4), and indeed was concomitant with expression of a novel *NODAL* exon defined by the predicted splice site in hES cell lines (Figure 2.6). Direct manipulation of the SNP allele in a minigene system provided experimental evidence that rs2231947 can directly affect the alternative splicing of *NODAL* (Figure 2.7).

Perhaps surprisingly, many reports of putatively alternatively spliced transcripts do not present minigene analyses or comment on the potential proximity of SNPs or other variation that may impact splice site selection. This is important given that allele specific splicing has been documented in human cells [38]. In the absence of a minigene analysis, the presence of two distinct transcripts from the same gene locus is not by itself sufficient evidence to demonstrate true alternative splicing from a single locus. It is of course possible that heterozygous SNP alleles could result in distinct transcripts being expressed from only one of two chromosomes in a diploid cell. The inclusion of genetically distinct minigenes, expression of two distinct *NODAL* isoforms in the rs2231947 homozygous T|T H9 hES cell line, and the presence of a single allele of a heterozygous SNP in both *NODAL* and *NODAL* variant transcripts collectively demonstrate true alternative splicing of *NODAL*.

In general, the *NODAL* variant transcript made up a small proportion of total *NODAL* transcript, despite the strong predicted strength of the alternative exon splice donor site. This suggests other genomic epigenetic elements likely discourage constitutive splicing of the alternative exon. This exemplifies the non-deterministic nature of splicing patterns, in that splice site motifs alone cannot perfectly predict splice site utilization or efficiency. Low *NODAL* variant transcript levels were even observed for the H9 cell line homozygous for the splice site-contributing T allele of rs2231947. I did not examine if there was an allelic dose effect on the proportion of *NODAL* variant spliced given variability in the *NODAL* splicing proportions between different isolates of the same cell line.

It is tempting to speculate how the associations and functional finding related to SNP rs2231947 may be mechanistically connected. An intriguing hypothesis is that the T allele for rs2231947 potentiates expression of the *NODAL* variant transcript isoform, which preferentially negatively affects the derivation of male hES cell lines. Of course, testing such a hypothesis would be difficult given the large number of cell lines assessed, the retrospective nature of the analysis, and the ethical considerations of research on human embryos. The fact that I have determined a function for a SNP with interesting associations is in no way sufficient to declare it functionally responsible for the observed associations. Others have emphasized that restraint should be exercised in assigning causality to polymorphisms based on associations, or even indirect experimental evidence [39]. Indeed, even the best and largest genetic association studies often fail to identify any functional polymorphisms at all, let alone the causal variant(s) responsible for the given trait of interest. This is undoubtedly due to the general inability to model endogenous complexity concerning how SNP alleles affect intricately controlled gene expression networks, as well as the difficulty in untangling contributions of individual SNPs within larger haplotypes.

Indeed, it cannot be ignored that genetic variations such as SNP alleles are inherited in the context of chromosomes, leading to varying degrees of high linkage disequilibrium (LD) between nearby SNPs that constitute a haplotype. Therefore, in any study of inherited genetic variation, it is informative to perform detailed LD analysis to define the

linkage group marked by the SNP of interest, as I have done here. Although such analysis is often overlooked, it has become increasingly appreciated in recent years as many SNPs identified through GWA studies have been found to be non-functional (at least to the extent to which their function can be assessed given how well current models recapitulate endogenous biology). The number of genomic variants catalogued by efforts such as the 1000 Genomes Project now makes it possible to derive a comprehensive list of genetic variants in high LD with a SNP of interest using reference populations of various ancestries. Furthermore, these SNP loci can be cross-referenced with numerous genomic annotations that have become increasingly available from large scale projects such as the ENCODE project. Together, these analyses allow for the identification of potentially functional variants for further study. Of course, SNPs in regions with no annotations may still be functional, as certainly many functional genomic elements remain to be discovered.

The relatively small linkage group with high LD to rs2231947 provided a manageable number of SNPs that may have contributed to the associations reported here. Several of these SNPs lie in experimentally validated *NODAL* enhancer regions. The most interesting of these was the proximal epiblast enhancer (PEE), where three SNPs in the rs2231947 linkage group are located. Since human ES cells have been shown to represent a similar developmental state to "primed" mouse epiblast stem cells (reviewed in [40]), it is possible the PEE identified in mouse epiblast cells is also an active driver of *NODAL* expression in hES cells. This is supported by the ENCODE data, as the *NODAL* PEE contains a DNase hypersensitivity site present all three hPSCs surveyed (H1 and H7 hES cells, and an iPS cell line). There is also a transcription factor "hotspot" within the PEE bound by 18 different transcription factors in H1-hES cells (Figure 2.9). No transcription factor binding sites were detected in this region in any non ES cell lines assayed so far, suggesting that this is an active regulatory region specific to ES cells. In the rs2231947 linkage group, SNP rs35210846 lies within 14 of these binding sites, and is found adjacent to (within 15 base pairs of) three more. These include binding sites for both NANOG and OCT4—two master regulators of stem cell pluripotency. Thus, SNP rs35210846 is a good candidate to also be functional in hES cells. One hypothesis is that SNP rs35210846 affects *NODAL* expression in hES cells through modulation of NANOG

and/ or OCT4 binding at the PEE. If this is the case, experimental modelling of this LD group would have to consider the combinatorial effects of alternative *NODAL* splicing as well as *NODAL* enhancer activity. I have fully characterized a linkage group of fourteen SNPs with two interesting associations in hES cells, identifying rs35210846 as a potentially functional SNP, and providing detailed experimental evidence for the *NODAL* splicing function of rs2231947.

The conservation analysis conducted here suggests that the splice donor site for the alternatively spliced *NODAL* exon is moderately conserved in mammals. Splice donor site motifs are generally highly conserved across vertebrates [18]. Therefore, it is possible that other species also contain functional splice sites for inclusion of a *NODAL* cassette alternative exon similar to the human exon described here. Of the 42 species with sequence alignment that included the human G[C/T] splice donor site dinucleotide, 4 have a GT dinucleotide matching the derived allele for human rs2231947. Since the human major allele (C) is the ancestral allele, the human SNP is not the first time that a putative canonical U2 splice site has evolved at this locus. It is also interesting that the ancestral C allele contributes to a non-canonical U2 GC-AG intron (canonical introns are defined by GT-AG). These non-canonical introns constitute approximately 0.7% of introns in both humans and mice [18]. There is evidence of evolutionary switching between these two U2 subtypes (either GT-AG to GC-AG or vice versa) occurring between species [18]. However, analysis of functional GC-AG splicing events reveals stricter conservation at surrounding bases compared to canonical splice sites, presumably to strengthen the weakened binding affinity of the C versus T nucleotide. Non canonical GC-AG introns have an average information content of 12.4 bits, while the canonical GT-AG splice sites have an information content of only 8.2 [18]. However, in the context of the human *NODAL* alternative exon splice site, the C allele was not predicted (Figure 2.4) or shown experimentally (Figure 2.7) to contribute to a splice site. It is possible that functional GC-AG splicing could take place in non-human species with adjacent sequence differences or other factors that contribute to enhanced binding affinity.

Evidence of sequence and element conservation within the cassette alternative exon locus also supports the possibility of conserved alternative splicing of *NODAL*. Lower base-

wise conservation scores for the alternative exon compared to the second constitutive exon were not surprising given *NODAL* coding regions are highly conserved; homozygous *Nodal* deletion is embryonic lethal in mouse [41, 42]. Relative to constitutively spliced exons, alternatively spliced exons are often recently evolved and not as well conserved between species. For example, in an analysis of alternatively spliced human exons, only 46% were found to be conserved in mouse, with only 7% of those found to also be alternatively spliced in mouse [43]. Notably, although the *NODAL* cassette alternative exon locus is somewhat conserved across mammals, the entire alternative exon locus (and almost the entire intron it lies within) shares no sequence alignment with mouse. Collectively, these analyses suggest there may be some conserved function to the alternatively spliced *NODAL* exon locus, although the novel *NODAL* transcript described here represents a major difference in *NODAL* gene expression between mouse and human stem cell models.

Population differentiation analysis suggests that individuals (and thus cell lines) of East Asian ancestry rarely have the minor allele for rs2231947, suggesting that the alternative *NODAL* isoform is not widely expressed in hES cell lines of these ancestries. One prediction of this low minor allele frequency is that hES cell lines of East Asian origin may not display the female sex bias. This analysis was not possible with the ISCI dataset due to low statistical power provided by a small number of such cell lines available for analysis. It is difficult to determine if this population differentiation is a result of drift, or negative selection in the East Asian population. Indeed, several human populations including the East Asian ancestral population endured a strong and sustained population bottleneck between 15,000 and 20,000 years ago [22], an event that can enhance both selection and genetic drift.

To the best of my knowledge, this work is the first to identify SNPs in association with any characteristics of human pluripotent cells. Thus, this study adds considerable and tangible depth to our understanding of inter-cell line heterogeneity in hES cells. The genetic associations reported here suggest specific genetic variation encompassing the *NODAL* gene as a source of previously reported and unexplained phenomena including the female sex bias in hESC lines [33] and highly variable XIST expression levels in

female hESC lines [5]. I also demonstrated how a single SNP in a non-coding region can have a large impact on gene expression. In this case, rs2231947 promotes the alternative splicing of a novel *NODAL* transcript that has never been described. The intronic SNP rs2231947 is likely one of many "buried treasure[s] within our genes" [44], that promote splicing of diverse transcript variants yet to be discovered [45]. Lastly, I have provided a comprehensive characterization of the genetic linkage group that it tags, in the process identifying another potentially functional SNP in hES cells. The H9 cell line represents a genetically rare sample for the linkage group marked by rs2231947, but has been extensively relied on as the primary model of early human embryonic development and pluripotency. This study is only one such manifestation of how genetics can both make substantial contributions to, and confound or bias, the study of biology.

2.4 Methods

2.4.1 Single nucleotide polymorphism (SNP) analysis

Human embryonic stem cell rs2231947 genotypes were obtained from dataset GSE33522 [5] from the NCBI's Gene Expression Omnibus (GEO). These data are the result of the largest effort to genotype the most frequently used human embryonic stem (hES) cell lines on a global scale. The sex, ancestry, and genetic relatedness of the cell lines were determined from supplementary files obtained from [5] and directly from the authors of the study as needed. For identical and related cell lines, the cell line with the lowest sample number was kept for analysis while all other related cell lines were excluded. Cell lines classified according to [5] as "European" or "Middle East-East European" were used for genotype analysis since they represented the largest cohort of cell lines with highly similar genetic ancestry according to principal component analysis performed in [5]. Cell line cohorts of other ancestries contained too few cell lines for independent genotype analysis. Genotypes were also obtained for rs2231947 from individuals included in the 1000 Genomes Project as a reference population for comparison. 1000 Genomes Project data was downloaded on September 2014 from the most current release (release 5 of phase 3). Two-tailed Fisher exact tests were performed using GraphPad software. Since I was specifically interested in SNP rs2231947 for its effect on *NODAL* splicing, this was the only SNP tested for any association and was not part of a genome

wide association study. Therefore, reported P values did not require correction for multiple hypothesis testing. Odds ratios were calculated by VassarStats (http://vassarstats.net/odds2x2.html). Forest plots were generated using the metafor package (http://www.metafor-project.org/doku.php) in R (https://www.r-project.org/).

All genotype, linkage disequilibrium, and population differentiation (F_{st}) analyses were performed using VCFtools version 0.1.12b (http://sourceforge.net/projects/vcftools/files/) and Samtools version 1.1 (http://sourceforge.net/projects/samtools/). LD analysis was conducted considering all polymorphisms within 1Mb (±500 kb) of rs2231947 (at hg19 chr10:72193854). All analyses were conducted with only genetically unrelated (no first, second, or third order relationship detected) founder individuals for all populations assessed.

For human pluripotent cell lines cultured in the lab, genomic DNA was isolated from hES and iPS cell lines using the Wizard Genomic DNA Purification Kit (Promega; Madison, Wisconsin, USA). SNP rs2231947 genotypes for cell lines not included in GEO dataset GSE33522 were determined using PCR amplification of the rs2231947 locus followed by restriction fragment length polymorphism (RFLP) assays specific to each of the C and T alleles.

For any cell line with known rs2231947 genotype from [5], XIST expression data was obtained from [4]. For two cell lines with expression reported for more than one sample, the mean of these expression values was used for analysis. ΔCt values were converted to fold changes relative to the median expression of cell lines with C|C genotype for rs2231947, using the equation: fold change = 2-(ΔΔCt), where ΔΔCt = ΔCt sample - ΔCt median C|C. An unpaired one-tailed t-test was performed using GraphPad software.

2.4.2 Splice site prediction analysis

Splice site prediction analysis was conducted using the "Automated Splice Site And Exon Definition Analyses" web server (splice.uwo.ca) and described in [46]. Sequence logos in Figure 2.4 were created using WebLogo version 2.8.2 (http://weblogo.berkeley.edu/ and [47, 48]).

2.4.3 Cell culture

Human Embryonic Stem (hES) cell lines and human induced Pluripotent Stem (iPS) cell lines were maintained on irradiated CF-1 Mouse Embryonic Fibroblasts (GlobalStem; Gaithersburg, Maryland, USA) with standard media composed of knockout DMEM/F12 (Life Technologies; Carlsbad, California, USA), 20% knockout serum replacement (Life Technologies), 1X non-essential amino acids (Life Technologies), 2 mM glutamine (Life Technologies), 0.1 mM 2-mercaptoethanol (Thermo Fisher Scientific; Waltham, Massachusetts, USA), and 4 ng/ml of basic fibroblast growth factor (Life Technologies). Cells were passaged mechanically and harvested only from feeder-free conditions that consisted of growth on a Geltrex matrix (Life Technologies) with defined mTeSR1 media (Stem Cell Technologies; Vancouver, British Columbia, Canada). H1, H9 and HES-2 lines were purchased from WiCell (Madison, Wisconsin, USA), the CA lines were provided by Dr. Cheryle Seguin (The University of Western Ontario), and the iPSC lines were provided by Dr. Bill Stanford (University of Ottawa). HEK 293 (ATCC; Manassas, Virginia, USA) cells were maintained in DMEM (Life Technologies) supplemented with 10% fetal bovine serum (Life Technologies). All cells were cultured at 37°C with 5% CO_2 in a humidified environment.

2.4.4 *NODAL* splicing analysis

Total RNA was isolated using PerfectPure RNA Cultured Cell Kit (5-PRIME; Hilden, Germany) and included on-column DNase treatment. Reverse transcription was performed using the High Capacity cDNA Reverse Transcription Kit (Life Technologies) to reverse transcribe 1 µg total RNA. *NODAL* was amplified from 1/20th of the cDNA product for semi-quantitative PCR analysis using AmpliTaq Gold 360 Master Mix (Life Technologies). *NODAL* RT-PCR in Figure 2.5 used a forward primer in exon 2: TGTGAGGGCGAGTGTCC, and reverse primer in exon 3: GAGGCACCCACATTCTTCCA. An annealing temperature of 60°C was used. The *NODAL* variant transcript was identified by gel purification of the longer and unexpected *NODAL* amplicon in Figure 2.4 using the QIAquick Gel Extraction Kit (Qiagen; Hilden, Germany) followed by cloning with the TOPO TA Cloning Kit for Sequencing (Life Technologies) and subsequent DNA sequencing.

Real time PCR in Figure 2.6 was performed using Power SYBR Master Mix (Life Technologies). 1/20th of the cDNA corresponding to 50 ng RNA was loaded in duplicate for detection and quantification of constitutive *NODAL* and *NODAL* variant. Constitutive *NODAL* forward primer: TACATCCAGAGTCTGCTG. Constitutive *NODAL* reverse primer: CCTTACTGGATTAGATGGTT. *NODAL* variant forward primer: CTGTTGGGGAGGAGTTTCA. *NODAL* variant reverse primer: AGGCTTGGCATGGAGGATA. Cloned PCR products were sequenced to confirm amplicon identity. The cloned products were also linearized, quantified, and diluted to various concentrations (copy number/μL). These standards were run alongside samples to obtain standard curves to estimate the number of copies of *NODAL* and *NODAL* variant transcripts detected in each cell line. Primer sets were checked for specificity using melt curve analysis. An annealing/extension temperature of 55° C was used.

2.4.5 Minigene analysis

A portion of *NODAL* from upstream of the 3' end of constitutive exon 2 to downstream of the 5' end of constitutive exon 3 (the final most 3' exon) was used for minigene analysis. This fragment was amplified from H9 gDNA (for rs2231947 = T) or HEK 293 gDNA (for rs2231947 = C) for splicing analysis using the forward primer: GGGCTCCTGGATCATCTACC, and the reverse primer: ACTCTGCCATTATCCACATAC. The same primers were used to amplify "already spliced" *NODAL* and corresponding *NODAL* variant control fragments from H9 cDNA. The forward and reverse primers also included restriction sites for ClaI and AgeI, respectively. The *NODAL* amplicon was then digested with ClaI and AgeI (New England Biolabs; Whitby, Ontario, Canada) for insertion into the FRE5 minigene plasmid backbone described in [20]. Ligation was performed with the Rapid DNA Dephos & Ligation Kit (Roche Applied Science; Penzberg, Germany). Site directed mutagenesis was also performed using the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent; Santa Clara, California, USA) to mutate SNP rs2231947 in the "H9" minigene from T to C (sense strand) using the primer ATGCCAAGCCTCAGGCGGGATTCAGGGTCTC (mutated base underlined). Minigene plasmid DNA was transfected into HEK 293 cells with Lipofectamine 2000

(Life Technologies) following the manufacturer's protocol. 72 hours after transfection, total RNA was isolated for RT-PCR analysis. Primers specific to the minigene plasmid backbone were used to avoid amplification of endogenous *NODAL*. Forward primer: CAAAGTGGAGGACCCAGTACC. Reverse primer: GCGCATGAACTCCTTGATGAC.

2.4.6 Allelic expression analysis

Amplicons containing SNP rs1904589 were amplified from CA1 hES cell cDNA using AmpliTaq Gold 360 Master Mix (Life Technologies). For analysis of total *NODAL*, the forward primer: CCCAGGTCACCTTTTCCTTGG and reverse primer: TGAGAGACTGAGGTGGATTGTC were used. For ddPCR analyses, the Taqman assay C___1853986_10 (Applied Biosystems; Foster City, California, USA) was used using standard cycling conditions.

2.4.7 SNP loci characterization

Functional genome annotations overlapping with SNP loci were obtained using the UCSC Genome Browser's Variant Annotation Integrator (https://genome.ucsc.edu/cgi-bin/hgVai) and extracted from the hg19 assembly. Conservation scores and sequence alignments were obtained with the UCSC Table Browser and Genome Browser as appropriate.

2.5 References

1. Närvä, E., Autio, R., Rahkonen, N., Kong, L., Harrison, N., Kitsberg, D., et al. (2010). High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity. *Nature Biotechnology*, *28*(4), 371–377. doi:10.1038/nbt.1615

2. Funk, W. D., Labat, I., Sampathkumar, J., Gourraud, P.-A., Oksenberg, J. R., Rosler, E., et al. (2012). Evaluating the genomic and sequence integrity of human ES cell lines; comparison to normal genomes. *Stem Cell Research*, *8*(2), 154–164. doi:10.1016/j.scr.2011.10.001

3. Laurent, L. C., Ulitsky, I., Slavin, I., Tran, H., Schork, A., Morey, R., et al. (2011). Dynamic Changes in the Copy Number of Pluripotency and Cell Proliferation Genes in Human ESCs and iPSCs during Reprogramming and Time in Culture,

8(1), 106–118. doi:10.1016/j.stem.2010.12.003

4. Adewumi, O., Aflatoonian, B., Ahrlund-Richter, L., Amit, M., Andrews, P. W., Beighton, G., et al. (2007). Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. *Nature Biotechnology*, *25*(7), 803–816. doi:10.1038/nbt1318

5. Amps, K., Andrews, P. W., Anyfantis, G., Armstrong, L., Avery, S., Baharvand, H., et al. (2011). Screening ethnically diverse human embryonic stem cells identifies a chromosome 20 minimal amplicon conferring growth advantage. *Nature Biotechnology*, *29*(12), 1132–1144. doi:10.1038/nbt.2051

6. Andrews, P. W., Benvenisty, N., McKay, R., Pera, M. F., Rossant, J., Semb, H., et al. (2005). The International Stem Cell Initiative: toward benchmarks for human embryonic stem cell research. *Nature Biotechnology*, *23*(7), 795–797. doi:10.1038/nbt0705-795

7. Mosher, J. T., Pemberton, T. J., Harter, K., Wang, C., Buzbas, E. O., Dvorak, P., et al. (2010). Lack of population diversity in commonly used human embryonic stem-cell lines. *The New England journal of medicine*, *362*(2), 183–185. doi:10.1056/NEJMc0910371

8. Laurent, L. C., Nievergelt, C. M., Lynch, C., Fakunle, E., Harness, J. V., Schmidt, U., et al. (2010). Restricted ethnic diversity in human embryonic stem cell lines. *Nature Methods*, *7*(1), 6–7. doi:10.1038/nmeth0110-06

9. Rouhani, F., Kumasaka, N., de Brito, M. C., Bradley, A., Vallier, L., & Gaffney, D. (2014). Genetic Background Drives Transcriptional Variation in Human Induced Pluripotent Stem Cells. *PLoS Genetics*, *10*(6), e1004432–11. doi:10.1371/journal.pgen.1004432

10. Burrows, C. K., Banovich, N. E., Pavlovic, B. J., Patterson, K., Gallego Romero, I., Pritchard, J. K., & Gilad, Y. (2016). Genetic Variation, Not Cell Type of Origin, Underlies the Majority of Identifiable Regulatory Differences in iPSCs. *PLoS Genetics*, *12*(1), e1005793–18. doi:10.1371/journal.pgen.1005793

11. Kyttälä, A., Moraghebi, R., Valensisi, C., Kettunen, J., Andrus, C., Pasumarthy, K. K., et al. (2016). Genetic Variability Overrides the Impact of Parental Cell Type and Determines iPSC Differentiation Potential. *Stem Cell Reports*, *6*(2), 200–212. doi:10.1016/j.stemcr.2015.12.009

12. Kajiwara, M., Aoi, T., Okita, K., Takahashi, R., Inoue, H., Takayama, N., et al. (2012). Donor-dependent variations in hepatic differentiation from human-induced pluripotent stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(31), 12538–12543. doi:10.1073/pnas.1209979109

13. Vallier, L., Mendjan, S., Brown, S., Chng, Z., Teo, A., Smithers, L. E., et al.

(2009). Activin/Nodal signalling maintains pluripotency by controlling Nanog expression. *Development*, *136*(8), 1339–1349. doi:10.1242/dev.033951

14. Xi, Q., Wang, Z., Zaromytidou, A.-I., Zhang, X. H. F., Chow-Tsang, L.-F., Liu, J. X., et al. (2011). A Poised Chromatin Platform for TGF-β Access to Master Regulators. *Cell*, *147*(7), 1511–1524. doi:10.1016/j.cell.2011.11.032

15. Bertero, A., Madrigal, P., Galli, A., Hubner, N. C., Moreno, I., Burks, D., et al. (2015). Activin/Nodal signaling and NANOG orchestrate human embryonic stem cell fate decisions by controlling the H3K4me3 chromatin mark. *Genes & Development*, *29*(7), 702–717. doi:10.1101/gad.255984.114

16. Roessler, E., Pei, W., Ouspenskaia, M. V., Karkera, J. D., Veléz, J. I., Banerjee-Basu, S., et al. (2009). Cumulative ligand activity of NODAL mutations and modifiers are linked to human heart defects and holoprosencephaly. *Molecular Genetics and Metabolism*, *98*(1-2), 225–234. doi:10.1016/j.ymgme.2009.05.005

17. Barakat, T. S., & Gribnau, J. (2010). X chromosome inactivation and embryonic stem cells. *Advances in experimental medicine and biology*, *695*(Chapter 10), 132–154. doi:10.1007/978-1-4419-7037-4_10

18. Abril, J. F., Castelo, R., & Guigó, R. (2005). Comparison of splice sites in mammals and chicken. *Genome research*, *15*(1), 111–119. doi:10.1101/gr.3108805

19. Mucaki, E. J., Shirley, B. C., & Rogan, P. K. (2013). Prediction of Mutant mRNA Splice Isoforms by Information Theory-Based Exon Definition. *Human Mutation*, n/a–n/a. doi:10.1002/humu.22277

20. Orengo, J. P., Bundman, D., & Cooper, T. A. (2006). A bichromatic fluorescent reporter for cell-based screens of alternative splicing. *Nucleic Acids Research*, *34*(22), e148–e148. doi:10.1093/nar/gkl967

21. Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, *518*(7539), 331–336. doi:10.1038/nature14222

22. Donnelly, P., Gabriel, S. B., Green, E. D., Hurles, M. E., Knoppers, B. M., Marth, G. T., et al. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. doi:10.1038/nature15393

23. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. doi:10.1093/bioinformatics/btr330

24. Hinrichs, A. S., Raney, B. J., Speir, M. L., Rhead, B., Casper, J., Karolchik, D., et al. (2016). UCSC Data Integrator and Variant Annotation Integrator.

Bioinformatics, *32*(9), 1430–1432. doi:10.1093/bioinformatics/btv766

25. Norris, D. P., & Robertson, E. J. (1999). Asymmetric and node-specific nodal expression patterns are controlled by two distinct cis-acting regulatory elements. *Genes & Development*, *13*(12), 1575–1588.

26. Vincent, S. D., Dunn, N. R., Hayashi, S., Norris, D. P., & Robertson, E. J. (2003). Cell fate decisions within the mouse organizer are governed by graded Nodal signals. *Genes & Development*, *17*(13), 1646–1662. doi:10.1101/gad.1100503

27. Papanayotou, C., Benhaddou, A., Camus, A., Perea-Gomez, A., Jouneau, A., Mezger, V., et al. (2014). A Novel Nodal Enhancer Dependent on Pluripotency Factors and Smad2/3 Signaling Conditions a Regulatory Switch During Epiblast Maturation. *PLoS Biology*, *12*(6), e1001890–14. doi:10.1371/journal.pbio.1001890

28. Pan, G., & Thomson, J. A. (2007). Nanog and transcriptional networks in embryonic stem cell pluripotency. *Cell Research*, *17*(1), 42–49. doi:10.1038/sj.cr.7310125

29. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, *20*(1), 110–121. doi:10.1101/gr.097857.109

30. Felsenstein, J., & Churchill, G. A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular biology and evolution*, *13*(1), 93–104.

31. Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, *15*(8), 1034–1050. doi:10.1101/gr.3715005

32. CHEN, A. E., EGLI, D., NIAKAN, K., DENG, J., AKUTSU, H., YAMAKI, M., et al. (2009). Optimal Timing of Inner Cell Mass Isolation Increases the Efficiency of Human Embryonic Stem Cell Derivation and Allows Generation of Sibling Cell Lines. *Cell Stem Cell*, *4*(2), 103–106. doi:10.1016/j.stem.2008.12.001

33. Ben-Yosef, D., Amit, A., Malcov, M., Frumkin, T., Ben-Yehudah, A., Eldar, I., et al. (2012). Female Sex Bias in Human Embryonic Stem Cell Lines. *Stem Cells and Development*, *21*(3), 363–372. doi:10.1089/scd.2011.0102

34. Muller, P., Rogers, K. W., Yu, S. R., Brand, M., & Schier, A. F. (2013). Morphogen transport. *Development*, *140*(8), 1621–1638. doi:10.1242/dev.083519

35. Onai, T., Yu, J.-K., Blitz, I. L., Cho, K. W. Y., & Holland, L. Z. (2010). Opposing Nodal/Vg1 and BMP signals mediate axial patterning in embryos of the basal chordate amphioxus. *Developmental Biology*, *344*(1), 377–389. doi:10.1016/j.ydbio.2010.05.016

36. Shen, Y., Matsuno, Y., Fouse, S. D., Rao, N., Root, S., Xu, R., et al. (2008). X-inactivation in female human embryonic stem cells is in a nonrandom pattern and prone to epigenetic alterations. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(12), 4709–4714.

37. Dvash, T., Lavon, N., & Fan, G. (2010). Variations of X Chromosome Inactivation Occur in Early Passages of Female Human Embryonic Stem Cells. *PLoS ONE*, *5*(6), e11330. doi:10.1371/journal.pone.0011330.s006

38. Nembaware, V., Lupindo, B., Schouest, K., Spillane, C., Scheffler, K., & Seoighe, C. (2008). Genome-wide survey of allele-specific splicing in humans. *BMC Genomics*, *9*(1), 265. doi:10.1186/1471-2164-9-265

39. Freedman, M. L., Monteiro, A. N. A., Gayther, S. A., Coetzee, G. A., Risch, A., Plass, C., et al. (2011). Principles for the post-GWAS functional characterization of cancer risk loci. *Nature genetics*, *43*(6), 513–518. doi:10.1038/ng.840

40. Wu, J., & Izpisua Belmonte, J. C. (2015). Dynamic Pluripotent Stem Cell States and Their Applications. *Cell Stem Cell*, *17*(5), 509–525. doi:10.1016/j.stem.2015.10.009

41. Conlon, F. L., Lyons, K. M., Takaesu, N., Barth, K. S., Kispert, A., Herrmann, B., & Robertson, E. J. (1994). A primary requirement for nodal in the formation and maintenance of the primitive streak in the mouse. *Development*, *120*(7), 1919–1928.

42. Zhou, X., Sasaki, H., Lowe, L., Hogan, B. L., & Kuehn, M. R. (1993). Nodal is a novel TGF-beta-like gene expressed in the mouse node during gastrulation. *Nature*, *361*(6412), 543–547. doi:10.1038/361543a0

43. Pan, Q., Bakowski, M. A., Morris, Q., Zhang, W., Frey, B. J., Hughes, T. R., & Blencowe, B. J. (2005). Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends in genetics : TIG*, *21*(2), 73–77. doi:10.1016/j.tig.2004.12.004

44. Cooper, D. N. (2010). Functional intronic polymorphisms: Buried treasure awaiting discovery within our genes. *Human genomics*, *4*(5), 284–288.

45. Lu, Z.-X., Jiang, P., & Xing, Y. (2012). Genetic variation of pre-mRNA alternative splicing in human populations. *Wiley interdisciplinary reviews. RNA*, *3*(4), 581–592. doi:10.1002/wrna.120

46. Nalla, V. K., & Rogan, P. K. (2005). Automated splicing mutation analysis by information theory. *Human Mutation*, *25*(4), 334–342. doi:10.1002/humu.20151

47. Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome research*, *14*(6), 1188–1190.

doi:10.1101/gr.849004

48. Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, *18*(20), 6097–6100.

Chapter 3

3 Characterization of human *NODAL* locus RNA variants

3.1 Introduction

Transcription is an important point of control over gene expression. However, there are numerous other points at which gene expression is controlled prior to mRNA translation into protein. These include, but are not limited to, mRNA splicing and the coupled processes of transcription termination and polyadenylation. The precise locations that mRNAs are transcribed from and processed at can vary for products of an individual gene locus, resulting in multiple distinct mRNA products (reviewed in [1]). For example, the use of alternative transcriptional start sites can modulate the 5' end of a transcript, affecting the nature of the 5' untranslated region (UTR) or even translational start site usage and the resulting translated protein product. A more complex system involves antisense transcription whereby two transcripts are expressed from the same genomic locus—one from each complementary strand of the genome. These "natural antisense transcripts" or "NATs" yield a pair of RNAs with sequence complementarity—the extent of which depends on the amount of overlap in their shared genomic locus. Unsurprisingly, this complementary nature of natural antisense transcripts often confers the ability of one transcript to regulate the expression (translation or otherwise) of its antisense counterpart (reviewed in [2, 3]).

Alternative splicing is perhaps the most well studied and most utilized process the cell exploits for generating expanded transcript diversity, with estimates suggesting as many as 95% of multi-exon human protein coding genes undergo alternative splicing [4]. While several different types of splicing choices are possible, exon skipping is the most frequently observed event; an alternatively spliced exon flanked by two constitutive exons is either spliced into processed transcripts, or spliced out after being passed over as intronic sequence. Alternative splicing is a major mechanism regulating tissue-specific gene expression [4]. In human embryonic stem cells, induced differentiation is accompanied by widespread changes in alternative splicing [5, 6]. Perhaps unsurprisingly, alternative splicing is frequently deregulated in cancer (reviewed in [7-

9]), and cancer cells can hijack stem cell alternative splicing programs to enhance the maintenance of cancer stem-like cells [10].

Beyond differential inclusion of exons in linear processed transcripts, a more exotic form of splicing produces circular RNAs through “back-splicing” of downstream 5’ splice donor sites that form junctions with upstream 3’ splice donor sites of either their own exon or upstream exons, resulting in completely closed circular RNA transcripts lacking free ends. Although often generated from protein coding pre-mRNAs, these transcripts are not generally protein-coding, but can act to regulate gene expression either at the level of transcription or post-transcriptionally through modulation of miRNA activity [11].

Similar to splicing, alternative polyadenylation of transcripts is also pervasive, with at least 70% of mammalian mRNAs undergoing alternative polyadenylation (APA) [12, 13]. Alternative polyadenylation sites in the 3’ UTR can modulate RNA stability, nuclear export, susceptibility to miRNA targeting, and translation (reviewed in [14]). Like alternative splicing, APA is also involved in cell fate decisions in early development [15, 16]. Remarkably, control over alternative splicing and alternative polyadenylation can be coupled [4]. However, it appears as if this link is limited to the definition of 3’ terminal exons and selection of intronic polyadenylation sites [17, 18].

While great strides have been made in appreciating the global complexity and diversity of gene expression at the RNA level, for any given gene of interest, many aspects of its expression likely remain undiscovered or not well characterized. Furthermore, analysis of alternative splicing events is often limited to specific exon junctions, with the full-length nature of the corresponding transcripts rarely assessed. Indeed, alternatively spliced full-length transcripts containing open reading frames are only beginning to be appreciated and catalogued on a genome-wide scale [19].

Nodal is no exception, as many studies concerning its expression focus on transcriptional regulation in mouse systems [20-23]. The alternatively spliced *NODAL* transcript reported in the previous chapter was the first alternatively processed *NODAL* transcript discovered, while many molecular details of the constitutively spliced human *NODAL*

transcript have not been directly studied. This chapter fully characterizes the *NODAL* splice variants identified in the previous chapter, in terms of their transcriptional start sites and sites of polyadenylation. I also identify several additional RNAs transcribed from the *NODAL* locus, including a natural antisense transcript and a circular RNA. I also develop and validate droplet digital PCR (ddPCR)-based methods to potentiate absolute quantitative comparison between distinct transcript isoforms. Collectively, this work provides a comprehensive and quantitative assessment of *NODAL* locus transcript expression in human embryonic stem cells and human cancer cell lines and samples of various origin. I show that full-length constitutive *NODAL* transcripts are expressed at low levels in cancer samples, and that comprehensive analysis of *NODAL* transcript diversity helps to explain previously mentioned discrepancies hindering the confident detection of *NODAL* transcripts [24].

3.2 Results

To characterize the newly discovered genetically-regulated and alternatively spliced *NODAL* transcript isoform, a set of several primers were first used to assess exon junctions formed with the novel alternative exon (Figure 3.1A,B). Relative to constitutive *NODAL*, the *NODAL* variant isoform contains a 116 base cassette exon between the second and third constitutive exons. I next examined if a putative *NODAL* variant open reading frame (ORF) delineated by the canonical *NODAL* start codon and an alternative exon-induced premature termination codon (PTC) in constitutive exon 3 was present in hES cells. This ORF consisting of constitutive exon 1, constitutive exon 2, the alternatively spliced exon, and part of constitutive exon 3, was successfully amplified from H9 hES cell cDNA (Figure 3.1C,D). Notably, this cDNA was generated using oligo dT primers to convert only polyA tail-containing mRNAs. Hence, the alternative *NODAL* exon is spliced into full-length processed *NODAL* transcripts.

Beyond the open reading frames of the two *NODAL* isoforms, I sought to determine the transcript termini for each isoform that define the 5’ untranslated region (UTR) upstream of each start codon and the 3’ UTR downstream of each stop codon. For processed mRNA transcripts, 3’ ends are marked by the start of a polyA tail approximately 15-30 nucleotides downstream of a polyadenylation signal (PAS) (reviewed in [25]). Sequence

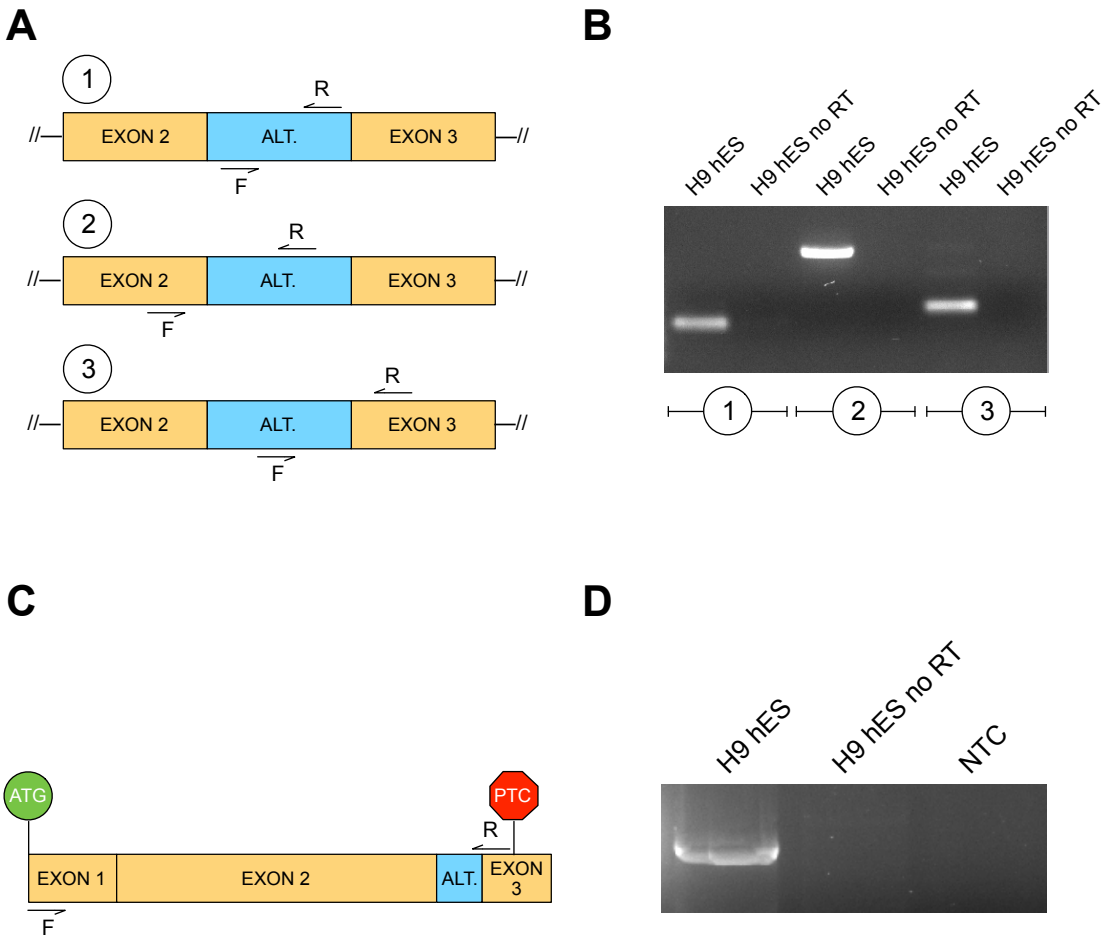


Figure 3.1: The alternative *NODAL* exon forms junctions with adjacent constitutive exons and is found within a fully spliced and polyadenylated open reading frame-containing transcript in H9 hES cells.
A) Locations of primers used in end-point PCR analysis in (B) of *NODAL* transcripts containing the cassette alternative exon. B) The alternatively spliced *NODAL* exon can be amplified from polyadenylated transcripts and forms junctions with constitutive exon 2 and constitutive exon 3 in hES cells. C) Primers used to target the predicted open reading frame contained within *NODAL* variant transcripts. D) The predicted *NODAL* variant open reading frame can be amplified from polyadenylated transcripts in hES cells. “F” = forward primer identical in sequence to the sense strand of *NODAL*. “R” = reverse primer of antisense sequence. “hES” = human embryonic stem cell. “RT” = reverse transcriptase. “NTC” = no template control. ATG marks the constitutive *NODAL* start codon. “PTC” = premature termination codon in constitutive exon 3 in frame with the *NODAL* variant reading frame. “ALT.” = cassette alternative *NODAL* exon.

analysis of *NODAL*'s constitutive exon 3 for common polyadenylation signals revealed two AUUAAA motifs and a single AAUAAA motif (Figure 3.2A). These two motifs are the most commonly utilized for polyadenylation of human transcripts [26], although other less-frequently utilized putative PASs were also found in the annotated 3' UTR. 3' rapid amplification of cDNA ends (3' RACE) for total *NODAL* transcript revealed two isoforms with distinct polyadenylation sites. Sequencing of these products confirmed that *NODAL* transcripts are alternatively polyadenylated in hES cells closely downstream of either a more proximal AUUAAA site, or a more distal AAUAAA site, at roughly equal levels (Figure 3.2A,B). Conducting the same procedure with primers designed to specifically detect *NODAL* variant transcripts also showed alternative usage of the same polyadenylation sites, but in a manner highly skewed toward the distal site (Figure 3.2C,D).

A similar approach to determine 5' ends of transcripts known as 5' RACE was conducted for total *NODAL* transcripts (Figure 3.3A), and specifically for the *NODAL* variant (Figure 3.3B). For total *NODAL*, a single product was obtained. Sequencing revealed a 5' end 14 bases upstream of the annotated *NODAL* translational start codon (Figure 3.3C), but 28 bases downstream of the annotated *NODAL* transcriptional start site in RefSeq. In contrast, several different products were detected for the *NODAL* variant (Figure 3.3D). The shortest and most abundant product corresponded to a 5' end within the coding region of constitutive exon 1. The middle band contained a more distal 5' end also within constitutive exon 1. Notably, it is possible that these products resulted from incomplete reverse transcription, and other samples did reveal 5' ends upstream of the *NODAL* translational start codon.

Surprisingly, the longest band did not contain any sequence from constitutive exon 1. Instead, there was a novel splice junction between constitutive exon 2 and a putative exon shortly upstream of constitutive exon 1. This product appeared to be enriched in the *NODAL* variant transcripts as no products containing this novel splice junction were detected in analysis of total *NODAL* transcripts.

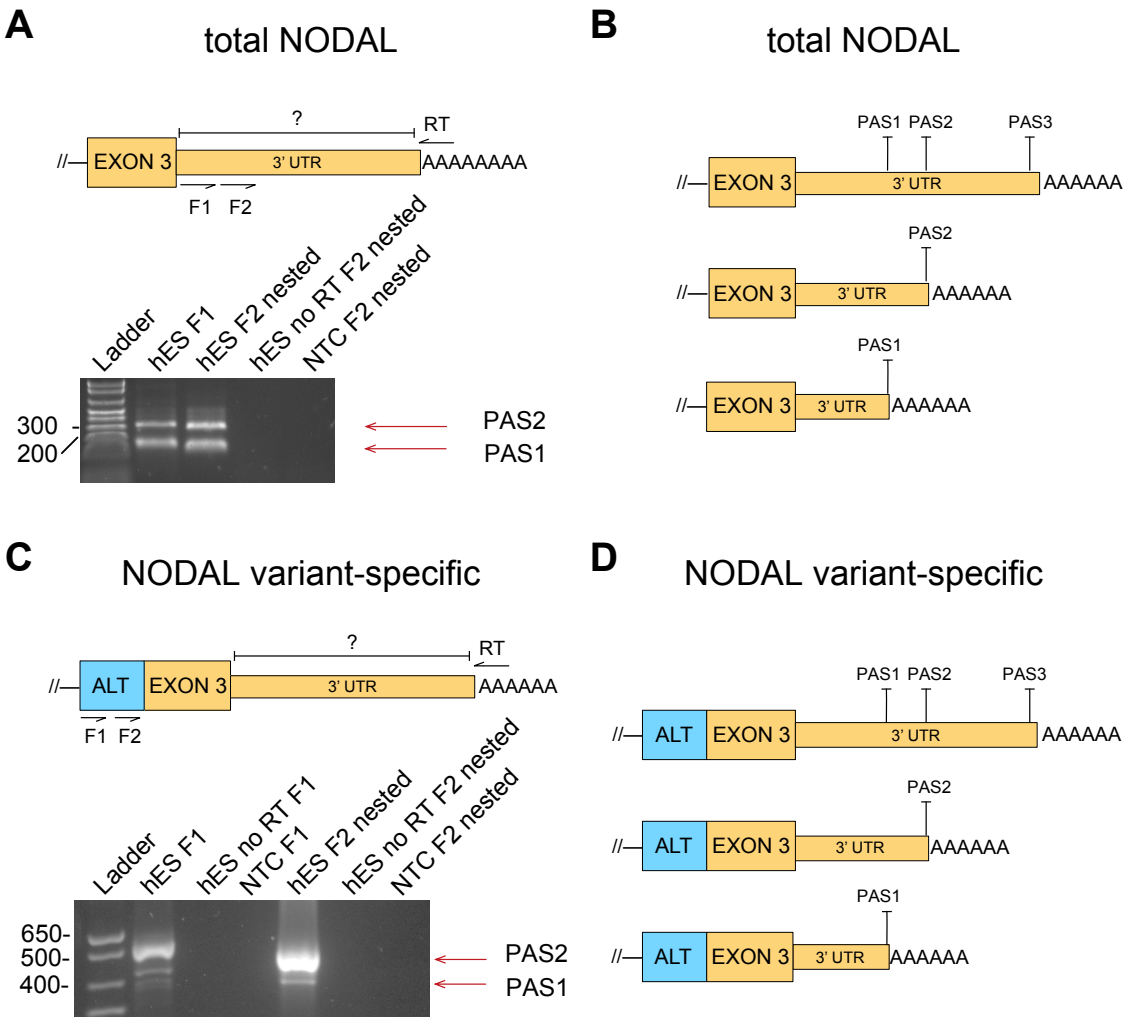


Figure 3.2: Both the *NODAL* variant and total *NODAL* are alternatively polyadenylated and utilize the same polyadenylation sites, but with different frequencies.

A) 3' transcript end analysis reveals roughly equal utilization of two *NODAL* polyadenylation sites. B) Sequencing mapped these sites to the two more proximal of three "canonical" polyadenylation sites defined by A[A/T]TAAA motifs in the annotated *NODAL* 3' UTR. C) and D) The same analysis limited to transcripts containing the cassette alternative exon reveals polyadenylation at the same two sites, but with usage skewed heavily toward the more distal site. "F1" and "F2" represents forward primers for initial PCR and nested PCR, respectively. "R" represents the reverse primer used to prime reverse transcription. "ALT" = cassette alternative exon. "UTR" = untranslated region. "PAS" = polyadenylation sequence. "AAAAAA" represents the polyA tail at the 3' end of transcripts. Numbers left of gels in A and C indicate size of DNA markers in base pairs. Exons upstream (5') of the alternative exon or exon 3 are not shown.

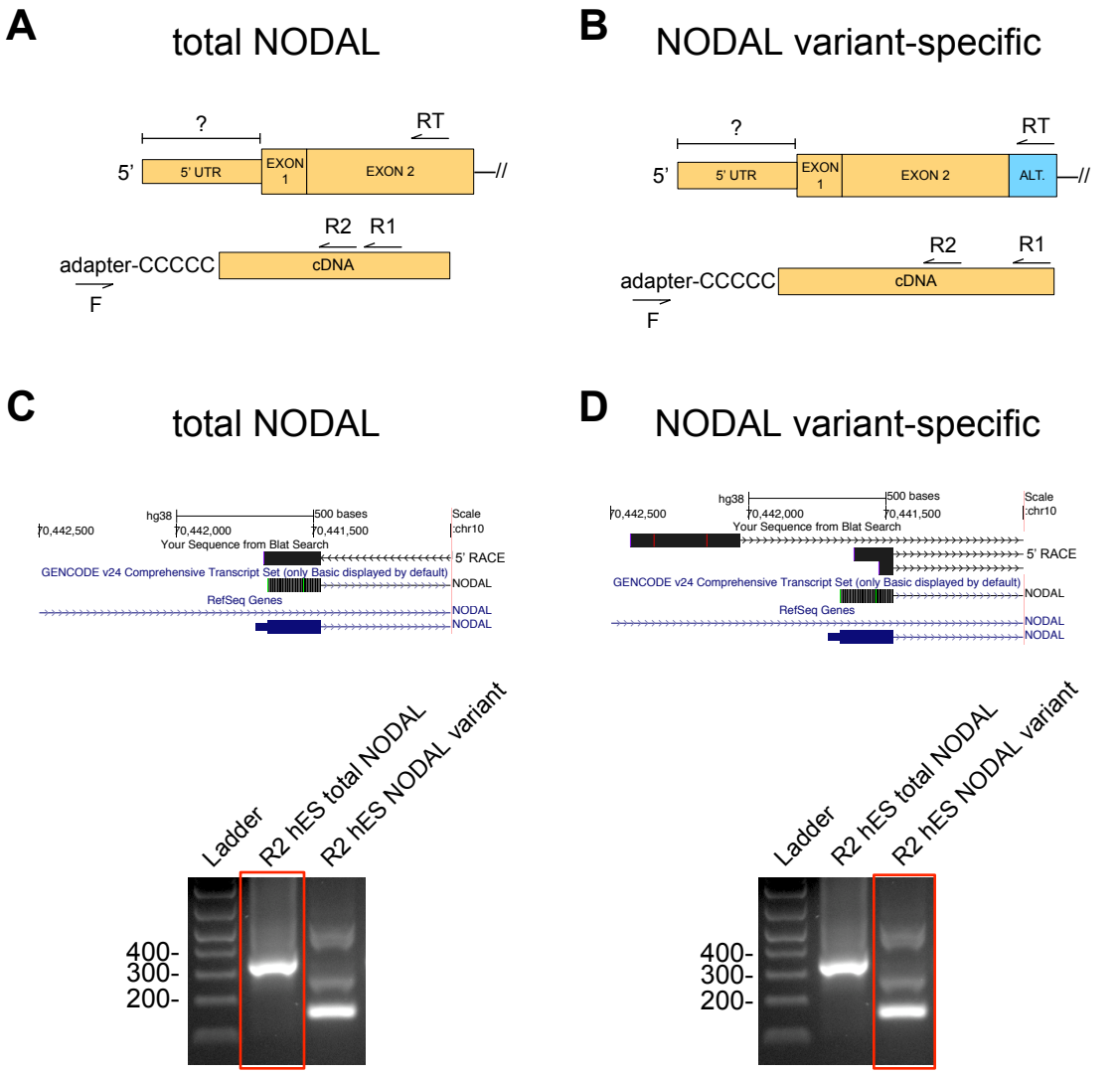


Figure 3.3: Discovery of an alternative 5' transcriptional start site and first exon of human *NODAL* that is enriched in *NODAL* variant transcripts relative to total *NODAL*.

5' RACE was conducted to determine the nature of the 5' ends of *NODAL* transcripts. A) Analysis of "total" *NODAL* transcripts was conducted with primers specific to constitutive exon 2. B) Analysis of *NODAL* variant transcripts was conducted with primers specific to the cassette alternative exon and constitutive exon 2. C) Only one distinct 5' end product was detected for total *NODAL*, corresponding to a short 5' UTR upstream of the annotated start codon. D) Several distinct products were obtained for analysis of *NODAL* variant transcripts. The shorter two products both had 5' ends mapping to constitutive exon 1 and likely resulted from incomplete reverse transcription. The longest of the three products did not contain any exon 1 sequence and revealed novel splicing to an alternative first exon upstream of constitutive exon 1. Numbers to the left of gels in C and D indicate size of DNA markers in base pairs.

Now that the full-length nature of constitutive and variant *NODAL* transcripts had been determined, in order to quantitatively study *NODAL* splicing, it was important to develop and validate detection assays that were both quantitative and isoform-specific. I developed a series of PCR assays to quantify each alternatively spliced *NODAL* isoform. The relative benefits and drawbacks of each of these methods are elaborated on in the Discussion.

The first and most commonly employed type of assay to assess exon skipping events is an end-point PCR assay that employs a single primer set; a forward primer targets a constitutive exon upstream of an alternative splicing event of interest, and a reverse primer targets a constitutive exon downstream [27]. The isoform ratio can be obtained by relative quantification of the two resultant bands after agarose gel electrophoresis. Another option is separate real-time PCR reactions; one for transcripts that include the alternative exon, and another to detect either both transcript variants or transcripts that skip the alternative exon [28]. An estimate of the isoform ratio is determined using standard curves of cloned or synthetic dsDNA corresponding to each splice variant of interest. Real-time PCR assays can be implemented using either non-specific fluorescent probes such as SYBR green, or sequence-specific fluorescent probes. Examples of end-point and real-time PCR assays developed for detection of human *NODAL* splice variants used in the previous chapter are shown in Figure 3.4.

In digital droplet PCR (ddPCR), a single sample is fragmented into approximately 20,000 droplets, each of about 1 nL in volume, prior to target amplification. This fragmentation allows a large number of physically isolated PCR reactions run in parallel, with many reactions containing zero or one copy of the target. This method offers absolute quantitation, increased sensitivity and precision relative to real time PCR assays, as well as detection at the level of single molecules. I next developed a duplexed ddPCR assay for simultaneous detection of alternatively spliced *NODAL* transcripts. This assay provided absolute quantification and completely specific detection of both constitutive *NODAL* and *NODAL* variant transcripts in a single assay (Figure 3.5A,B). A single probe ddPCR assay targeting the boundary between constitutive exon 1 and constitutive exon 2 was also effective in detection of total *NODAL* transcripts (Figure 3.5C,D).

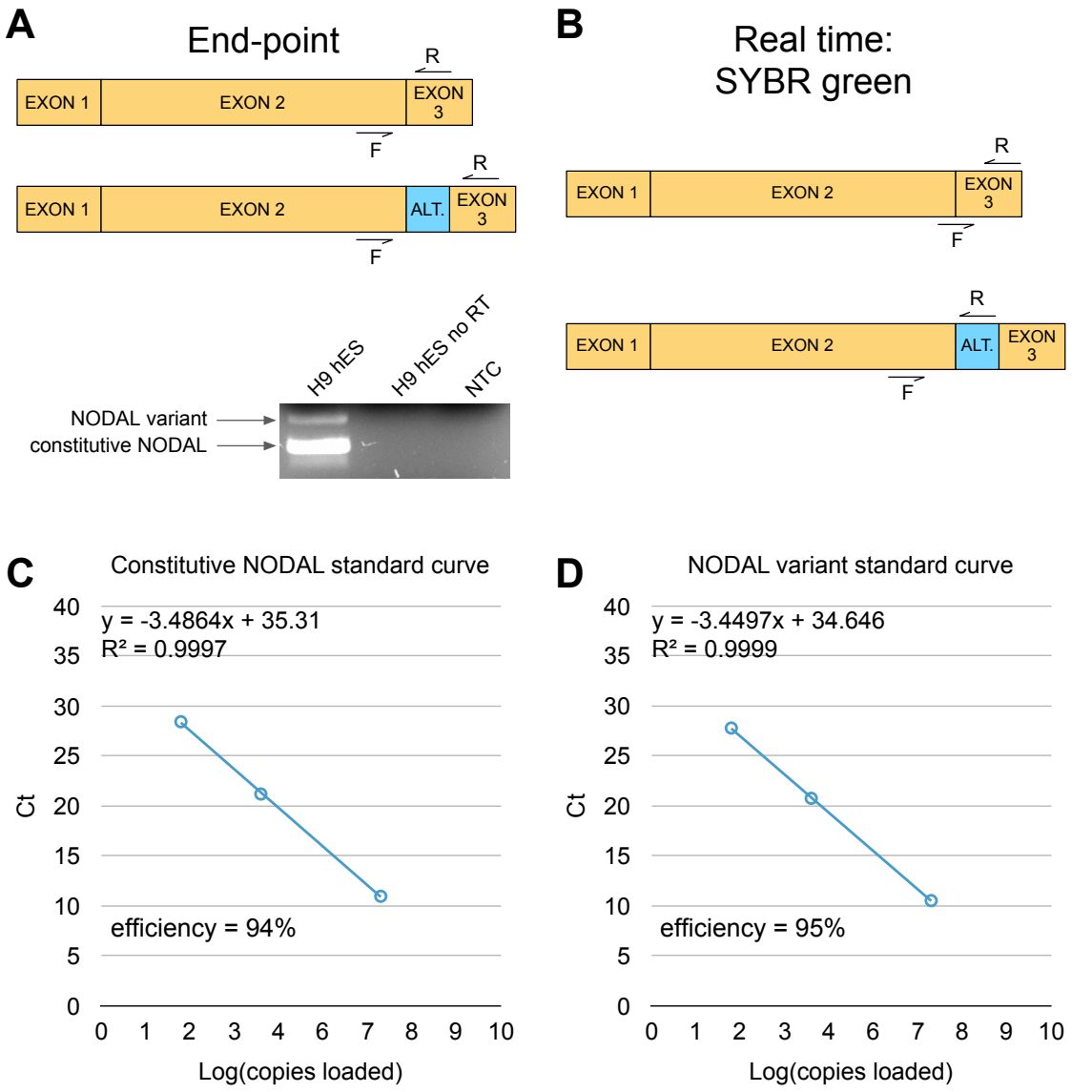


Figure 3.4: End-point PCR and real-time PCR assays for quantitative analysis of *NODAL* splice variant ratios.

A) Top shows a common strategy for the quantification of transcript isoforms resulting from alternative splicing. Bottom shows an example of human *NODAL* splicing analysis from chapter 2. B) An example of a real time PCR strategy for separate detection of *NODAL* variant and constitutive (or total *NODAL*) transcripts. C) and D) example standard curves for determination of constitutive *NODAL* (C) and *NODAL* variant (D) transcripts. Equations for standard curves and corresponding coefficients of determination (R^2) and amplification efficiencies are shown. “F” = forward primer (sense strand sequence). “R” = reverse primer (antisense strand sequence). “ALT.” = cassette alternative exon. “hES” = human embryonic stem cell. “RT” = reverse transcriptase. “Ct” = threshold cycle.

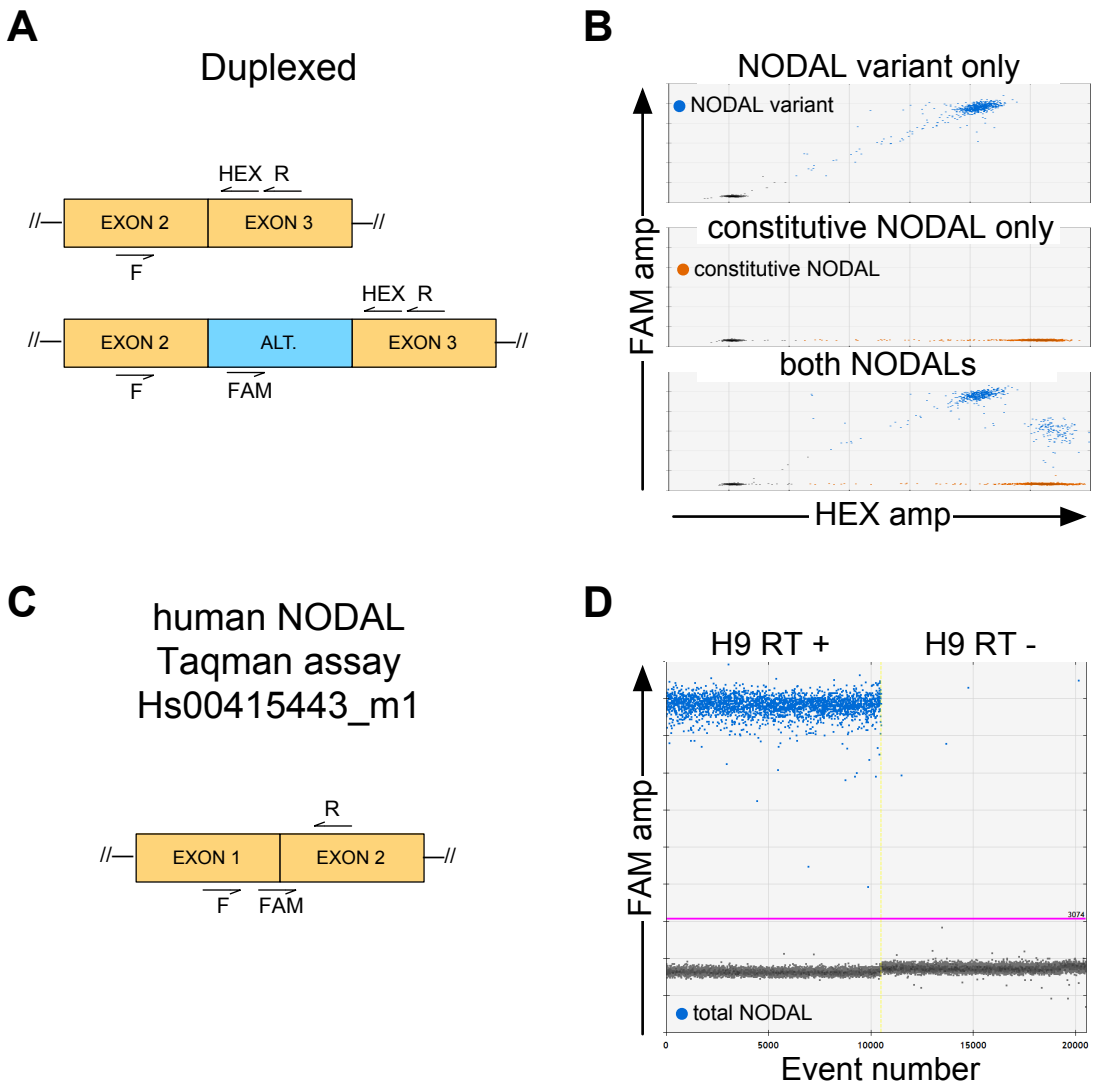


Figure 3.5: Droplet digital PCR assays for detection of *NODAL* splice variants and total *NODAL* transcript.

A) Design strategy for duplexed detection of *NODAL* splice variants using a HEX probe targeting common constitutive exon 3, and a FAM probe targeting the cassette alternative exon. B) Validation of the assay for specific detection of both constitutive *NODAL* and *NODAL* variant. C) Primer and probe layout for Taqman *NODAL* assay Hs00415443_m1. D) Validation of Hs00415443_m1 in ddPCR with H9 hES cells. “F” = forward primer (sense strand sequence). “R” = reverse primer (antisense strand sequence). “ALT.” = cassette alternative exon. “hES” = human embryonic stem cell. “RT” = reverse transcriptase.

In using these assays to detect *NODAL* expression levels, I noticed that *NODAL* was not reliably expressed in H9 hES cells, as transcript levels were extremely variable between samples of different passage cultured at different times in different locations. As an example, one such pair of samples differed in total *NODAL* expression by 3,000-fold, with only 26 copies of total *NODAL* transcript detected for the low-expressing sample in cDNA from 100 ng of total RNA. Notably, both samples expressed markers of pluripotency. (Figure 3.6A). Similarly, there was also variability in the ratio of *NODAL* variant to total *NODAL* transcript between hES samples of different passage, and this variability was evident even between cells of subsequent passage cultured under the same conditions. As an example, a second pair of samples differed in *NODAL* isoform ratio by five-fold (Figure 3.6B).

To experimentally investigate possible factors that may influence *NODAL* transcript levels, I focused on hES media, as either defined media such as mTESR-1, or media conditioned by mouse embryonic fibroblasts (MEFs), are regularly employed in the maintenance of hES cells. H9 hES cells adapted for culture on a Matrigel matrix in defined media were manually passaged. Half of the cells were kept in defined media (mTESR-1), while the other half were switched to MEF-conditioned media (see methods). When cells were ready to again be passaged, they were harvested for RNA. While the cells grown in defined conditions expressed low levels of total *NODAL* transcript, after only several days of culture in MEF-conditioned media, H9 hES cells displayed markedly increased *NODAL* transcript levels (Figure 3.7A). Furthermore, this effect could be reversed. After two more continuous passages in MEF-conditioned media, cells were again returned to defined conditions, and *NODAL* expression decreased by approximately the same factor as it had increased previously (Figure 3.7B). Therefore, the culture media system employed was identified as a factor that directly affected *NODAL* transcript expression in hES cells. Notably, cells expressed similar levels of markers of pluripotency (Figure 3.7C) and had morphology typical of pluripotent stem cells under both conditions (Figure 3.7D).

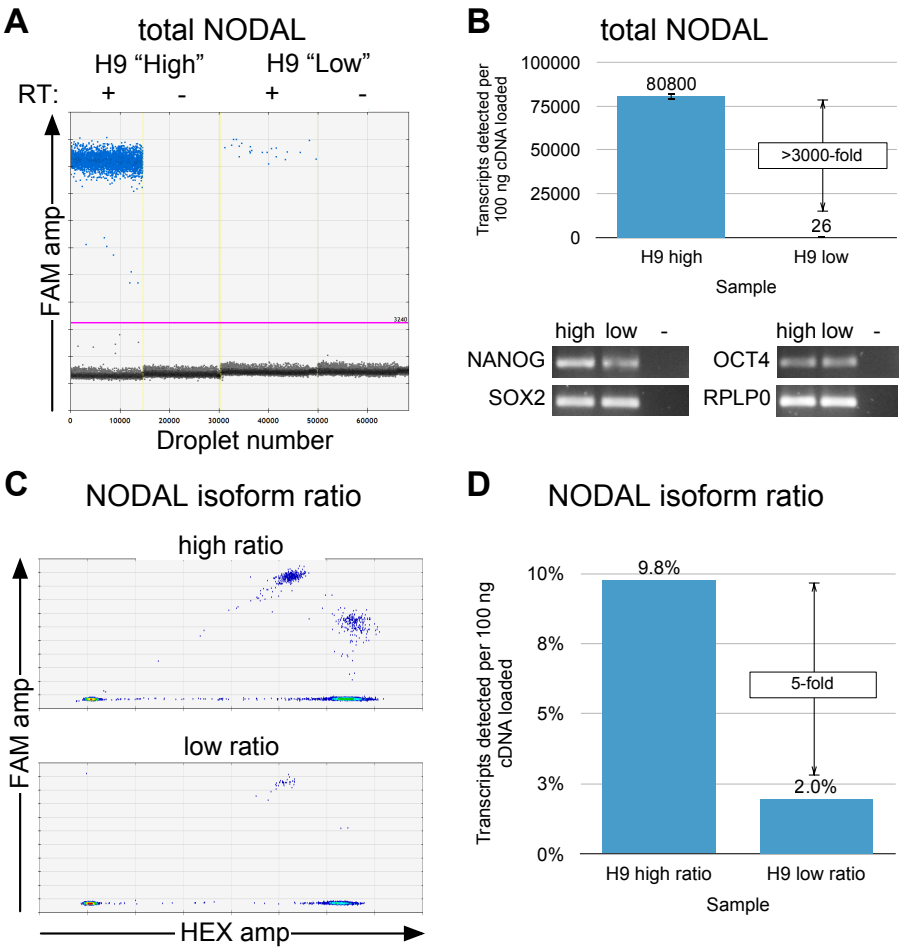


Figure 3.6: Highly variable expression and splicing ratio for *NODAL* transcript in H9 human embryonic stem cells.

NODAL expression can vary dramatically between RNA isolated at different times from cells grown under different conditions in different physical locations. A) ddPCR droplet plots for examples of "high" and "low" total *NODAL* expression in H9 hES cells using a primer probe assay spanning exon 1 and exon 2. "RT" = reverse transcriptase. Blue droplets are positive for *NODAL*, black droplets are negative. Pink line indicates arbitrary amplitude threshold for a positive call. B) Top: Total *NODAL* levels are more than 3,000-fold different between the "high" and "low" H9 hES samples. Error bars indicate 95% confidence interval for Poisson-calculated copies of transcript detected. Bottom: Both "high" and "low" *NODAL* samples were positive for markers of pluripotency using end-point RT PCR. "-" = no template control. C) 2D duplexed ddPCR plots for examples of "high" and "low" *NODAL* splice variant ratio (alternative exon included/ total *NODAL*) H9 hES samples. Heat map droplet view is shown. D) The *NODAL* isoform ratio is 5-fold different between the "high" and "low" *NODAL* isoform ratio samples.

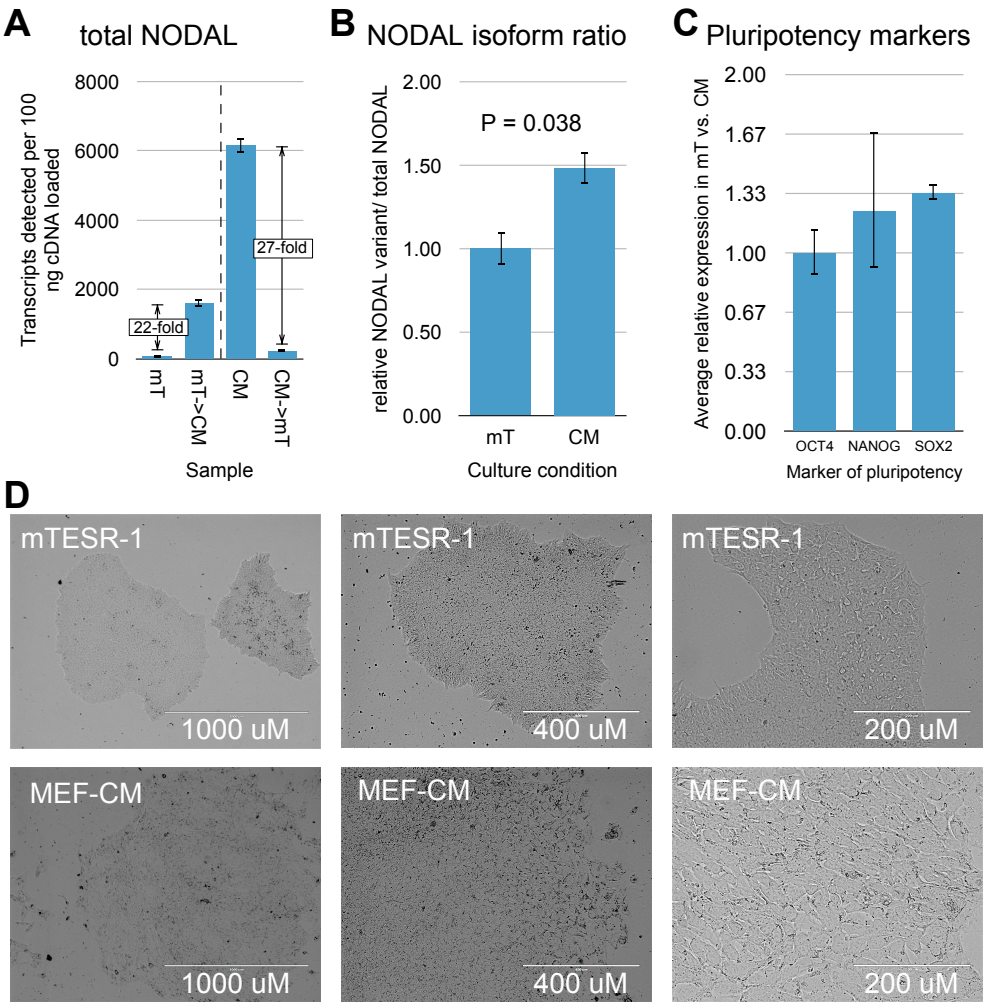


Figure 3.7: Total *NODAL* transcript levels and the proportion of alternatively spliced *NODAL* variant transcript are both reduced in H9 hES cells cultured in mTESR1 relative to MEF-CM.

A) H9 hES cells previously adapted to culture in defined mTESR-1 media expressed low levels of *NODAL* which were increased upon culture in MEF-CM. Subsequent return to mTESR-1 resulted in a reduction in *NODAL* levels of a similar magnitude. Error bars indicate 95% confidence interval for Poisson-calculated copies of transcript detected. B) The *NODAL* isoform ratio (*NODAL* variant/ total *NODAL*) was 50% higher for cells grown in MEF-CM relative to mTESR-1. The average of two samples from cells cultured in each media is shown. Error bars indicate standard deviations. P value is the result of a t-test. “mT” = mTESR-1. “CM” = mouse embryonic fibroblast-conditioned media. C) Expression of pluripotency markers was not lower in mTESR-1 relative to MEF-CM culture conditions. Error bars indicate standard deviations. D) Representative images of hES cells cultured in mTESR-1 (top) and MEF-CM (bottom) at increasing magnifications (left to right). Scale bars are shown.

Since aberrant expression of *NODAL* in numerous patient samples and human cancer cell lines has been described [29, 30], I next surveyed several human cancer cell lines of various origins for their levels of *NODAL* transcript. A survey of commonly utilized breast cancer cell lines that have previously been used to model *NODAL* biology showed variable but low expression of total *NODAL* transcript (Figure 3.8A). Notably, only 2 copies per 100 ng input RNA were detected for the triple-negative MDA-MB-231 cell line which has previously been used as a model where knockdown of *NODAL* reduces pro-tumourigenic phenotypes [31, 32], as has the included C8161 melanoma cell line [33], for which extremely low levels of *NODAL* transcript were also detected.

I was interested in investigating whether this unexpectedly low expression in cancer cell lines was a technical issue. To this end, two different MDA-MB-231 RNA samples isolated separately from different cell stocks were compared. Both samples revealed similar low levels of total *NODAL* transcript (Figure 3.9A). Next, a thermostable reverse transcriptase was utilized to determine if performing the reverse transcription reaction at an increased temperature improved reverse transcription efficiency through the partial denaturing of presumably complex secondary structure. When either random primers or oligo dT was used to prime thermostable reverse transcription, extremely low or undetectable levels of total *NODAL* transcript were still observed (Figure 3.9B). Low *NODAL* detection was also not limited to one specific assay, as transcript levels were not higher when using a primer probe assay targeting the exon 2 - exon 3 boundary (Figure 3.9B). Similarly, the use of a thermostable reverse transcriptase did not result in higher *NODAL* transcript detection in the H9 RNA sample with low *NODAL* expression (Figure 3.9C). Low *NODAL* levels were also detected even when cancer cell line samples were processed in parallel with high *NODAL*-expressing hES samples, and high levels of housekeeping genes such as *RPLP0* were detected in all samples. In summary, even the cancer cell line with the highest detected *NODAL* expression expressed more than 1,800-fold less transcript than hES samples with “high” *NODAL* expression, and several cell lines had no detectable *NODAL* transcript.

Since *NODAL* expression was low in many samples, including from H9 hES cells, I was interested in comparing assays that targeted different regions of the full-length transcript to determine if there were locus-dependent differential reverse transcription efficiencies

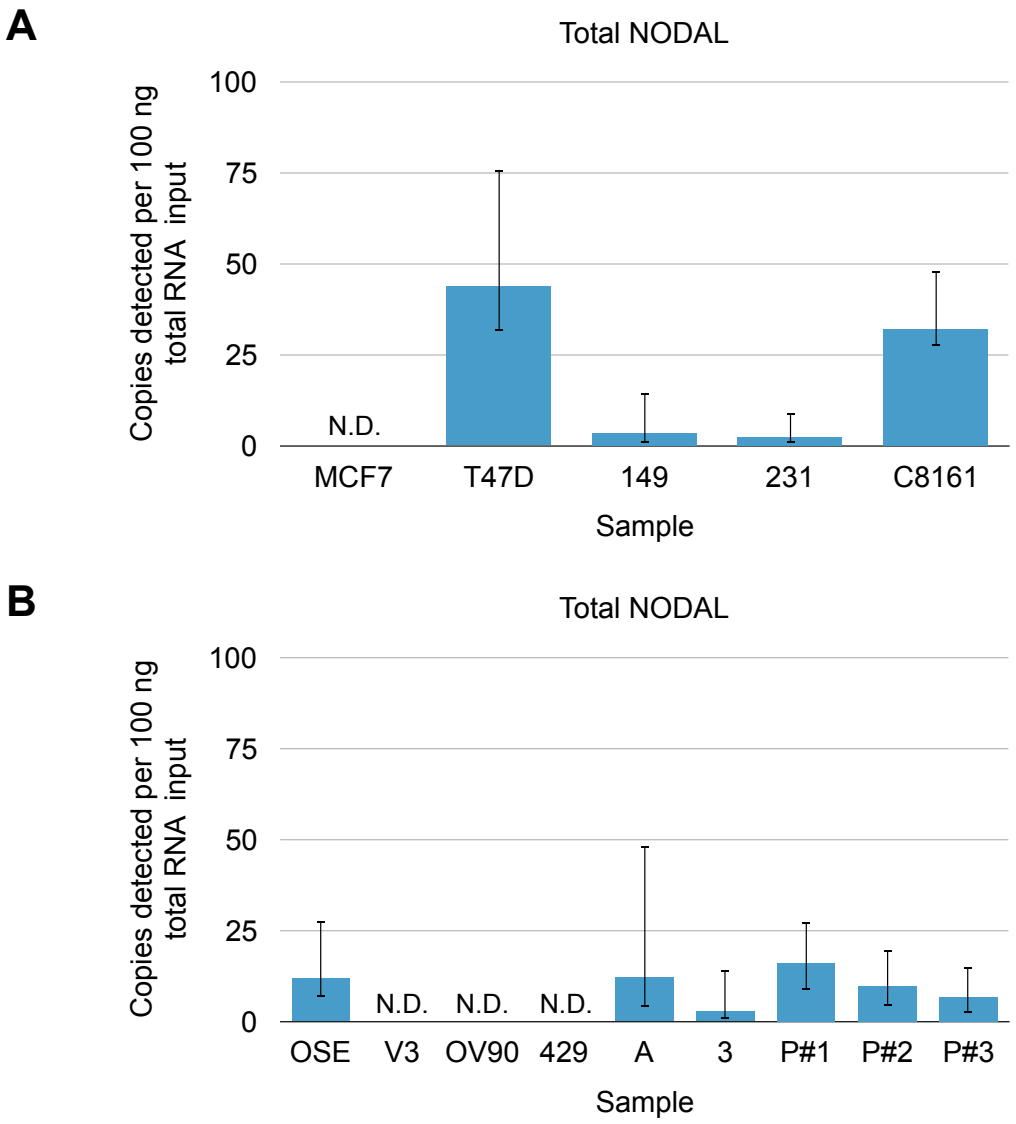


Figure 3.8: Quantitative analysis of total *NODAL* transcript levels reveals extremely low transcript abundance in human cancer cell lines and patient samples of various origin.

A) *NODAL* transcript levels were profiled in several human breast cancer cell lines of various subtypes and the C8161 melanoma line, most of which have been previously used to model *NODAL* function in cancer. “149” = SUM 149. “231” = MDA-MB-231. B) *NODAL* transcript levels were profiled in a panel consisting of one immortalized ovarian surface epithelial cell line (“OSE”), several ovarian carcinoma cell lines (“V3” = SKOV3, “429” = OVCA429, “A” = A2780S, “3” = OVCAR3), and three samples of carcinoma cells from patients with ovarian carcinoma, briefly cultured in vitro (P#1-3). “N.D.” = no transcript detected in cDNA from 100 ng total RNA input. Error bars indicate 95% confidence interval for Poisson-calculated copies of transcript detected.

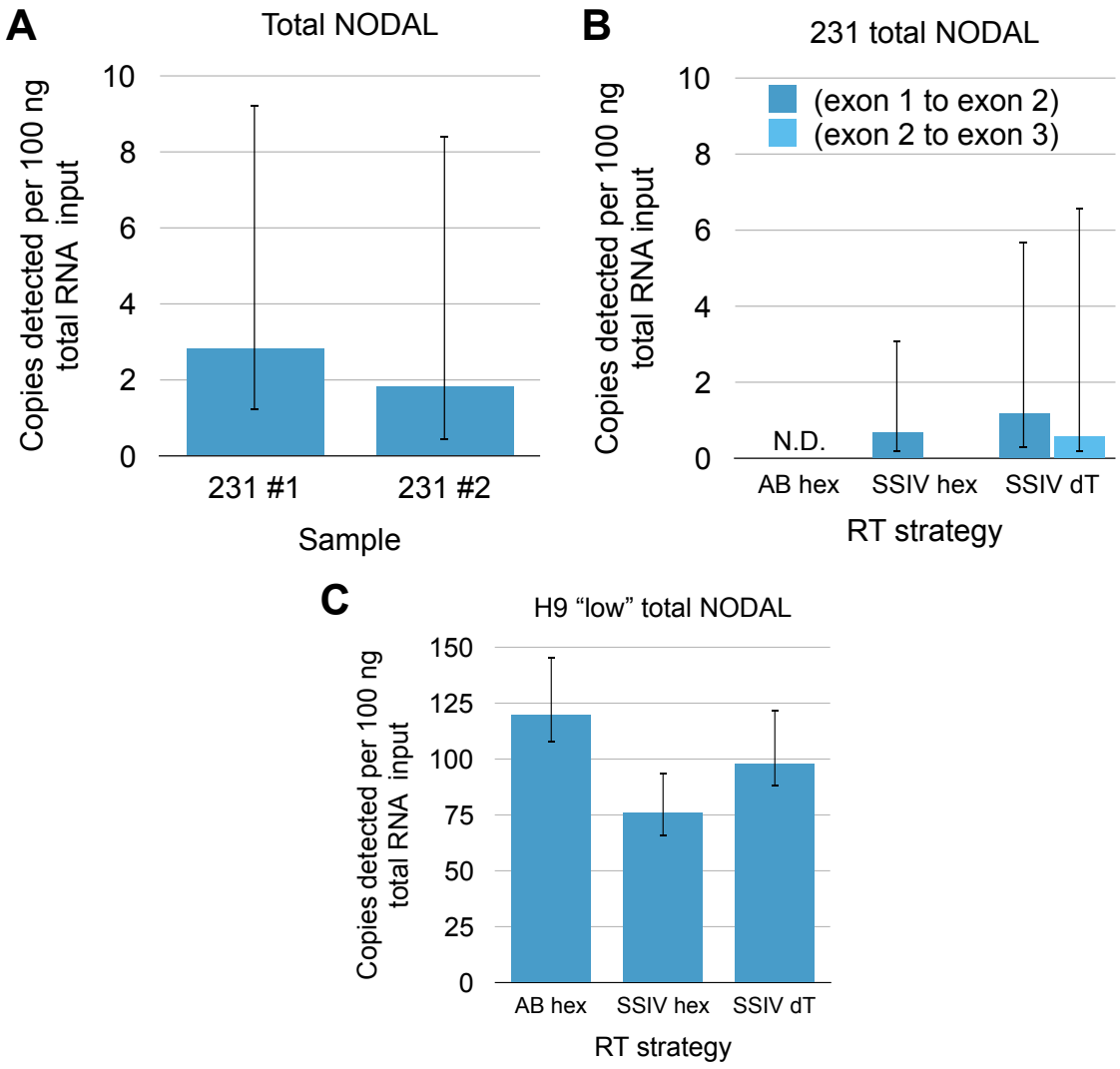


Figure 3.9: low *NODAL* expression is consistent and not improved by utilization of different reverse transcription strategies and PCR assays. A) MDA-MB-231 cells from different cell stocks both revealed virtually no detectable *NODAL* transcript. B) *NODAL* transcript levels remained extremely low or undetectable for a MDA-MB-231 sample from (A) when using different reverse transcription strategies or different primer probe assays to target different regions of the *NODAL* transcript. C) The use of oligo dT primers or thermostable reverse transcriptase also did not improve *NODAL* transcript detection in an H9 hES sample with "low" *NODAL* levels. "AB hex" = applied biosystems high capacity cDNA kit with RT primed by random hexamers. "SSIV hex" = SuperScript IV Reverse Transcriptase with RT primed by random primers. "SSIV dT" = SuperScript IV Reverse Transcriptase with RT primed by oligo dT. "RT" = reverse transcription. Error bars indicate 95% confidence interval for Poisson-calculated copies of transcript detected.

that would help improve detection. Such inter-assay comparisons are made possible by ddPCR since the method is absolutely quantitative and PCR amplification efficiency only influences the fluorescent magnitude of droplets and not the number of droplets positive for amplification used for quantification, thus the digital nature of the signal. In the H9 sample with "high" *NODAL* expression, the number of transcripts detected did not differ substantially between the assays: There was no more than a 1.5-fold difference seen when assays targeting the exon 2 - exon 3 junction, exon 1 – exon 2 junction, or exon 2 alone, were compared (Figure 3.10A). However, when the same assays were applied to the H9 sample with "low" *NODAL* expression, *NODAL* levels were 39-fold or 140-fold higher for the assay specific to exon 2, relative to the assays targeting exon 2 - exon 3, or exon 1- exon 2, respectively (Figure 3.10A). Since this assay did not cross an exon-exon junction, the inclusion of no reverse transcription controls demonstrated this signal was specific to RNA and did not result from genomic DNA or other DNA contamination of the RNA sample (Figure 3.10B). Collectively, these results suggest that increased signal in exon 2 is not the result of more efficient reverse transcription, and that an additional transcript sharing sequence with exon 2 may exist. A survey of human RNAs from Genbank [34] revealed AK001176 as a transcript that completely encompasses exon 2 of *NODAL*, extending about 500 bases upstream and downstream. Using primers internal to this transcript's annotated termini but both outside of *NODAL*'s exon 2 (within the adjacent introns), a product was detected from oligo dT reverse-transcribed RNA from hES cells (Figure 3.11A-B). Next, a ddPCR primer probe assay was developed that is specific to the AK001176 transcript but unable to detect constitutive exon 2 of *NODAL* (Figure 3.11C-D).

To verify the transcribed strand and orientation for AK001176, 3' RACE was conducted (Figure 3.12A). Similar to *NODAL*, two distinct products were detected. Sequences corresponding to the larger band revealed polyadenylation at a distal site about 200 bases downstream of a more proximal polyadenylation site corresponding to the smaller band (Figure 3.12B). While the proximal site revealed polyadenylation adjacent to a PAS with sequence "AGUAAA," the distal site was not proximal to any known PAS and was adjacent to a short polyA tract. The "AGUAAA" PAS was previously found to be the fourth most utilized PAS by human transcripts, although it should be noted that 15% of

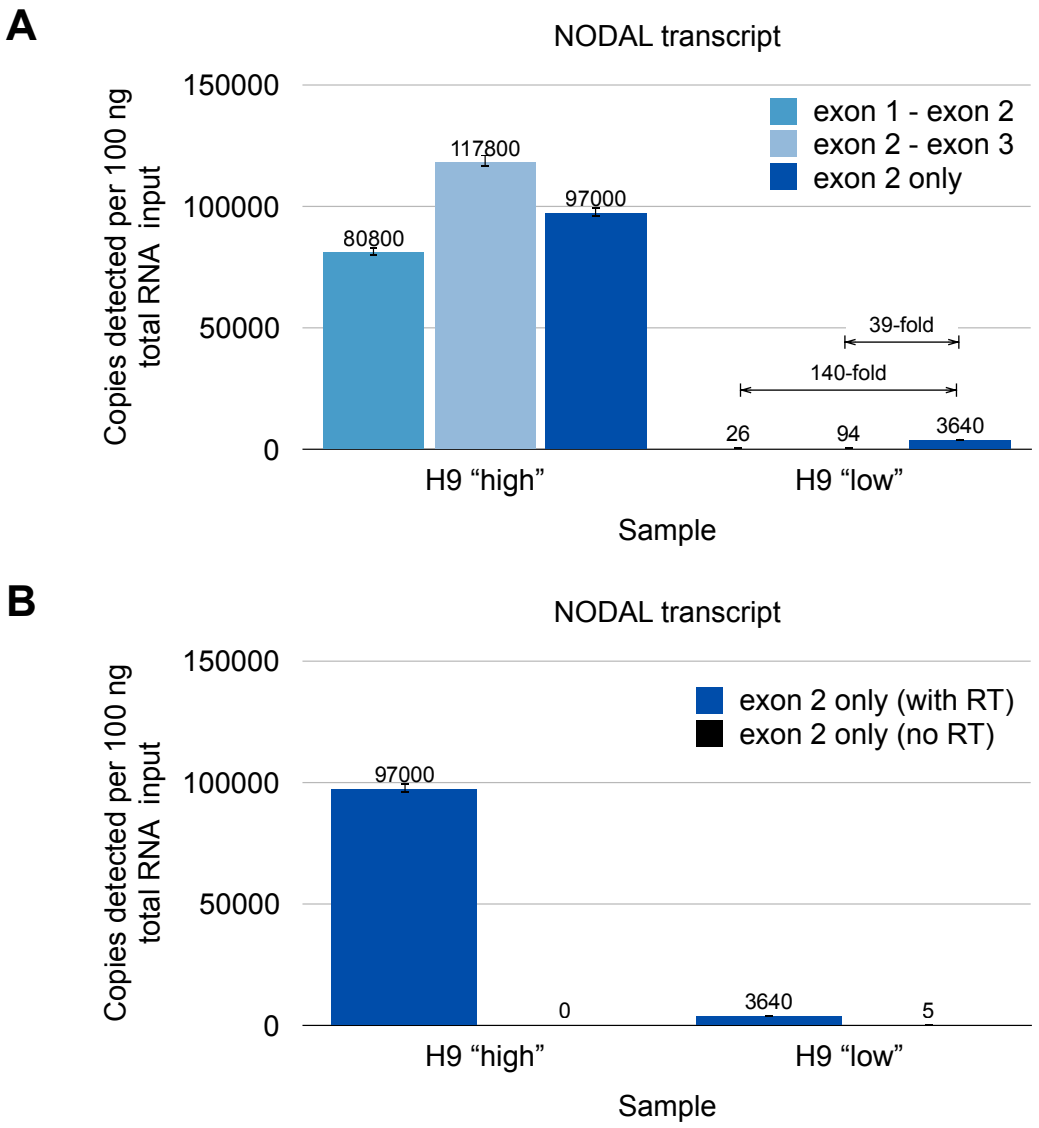


Figure 3.10: Higher levels of *NODAL* transcript were detected using a primer probe assay within constitutive exon 2.
A) For an H9 hES sample with high levels of *NODAL* transcript, similar levels (within 1.5-fold) were detected using assays targeting various regions of the transcript. Notably, utilizing the assay within constitutive exon 2 did not result in the highest levels of *NODAL* detection in this sample. In a sample with low *NODAL* transcript levels, the exon 2 assay detected over 38-fold more transcript than the next highest assay. B) Despite not crossing an exon-exon boundary, signal from the exon 2 assay did not result from genomic DNA contamination as no reverse transcriptase controls were negative. “RT” = reverse transcriptase. Error bars indicate 95% confidence interval for Poisson-calculated copies of transcript detected.

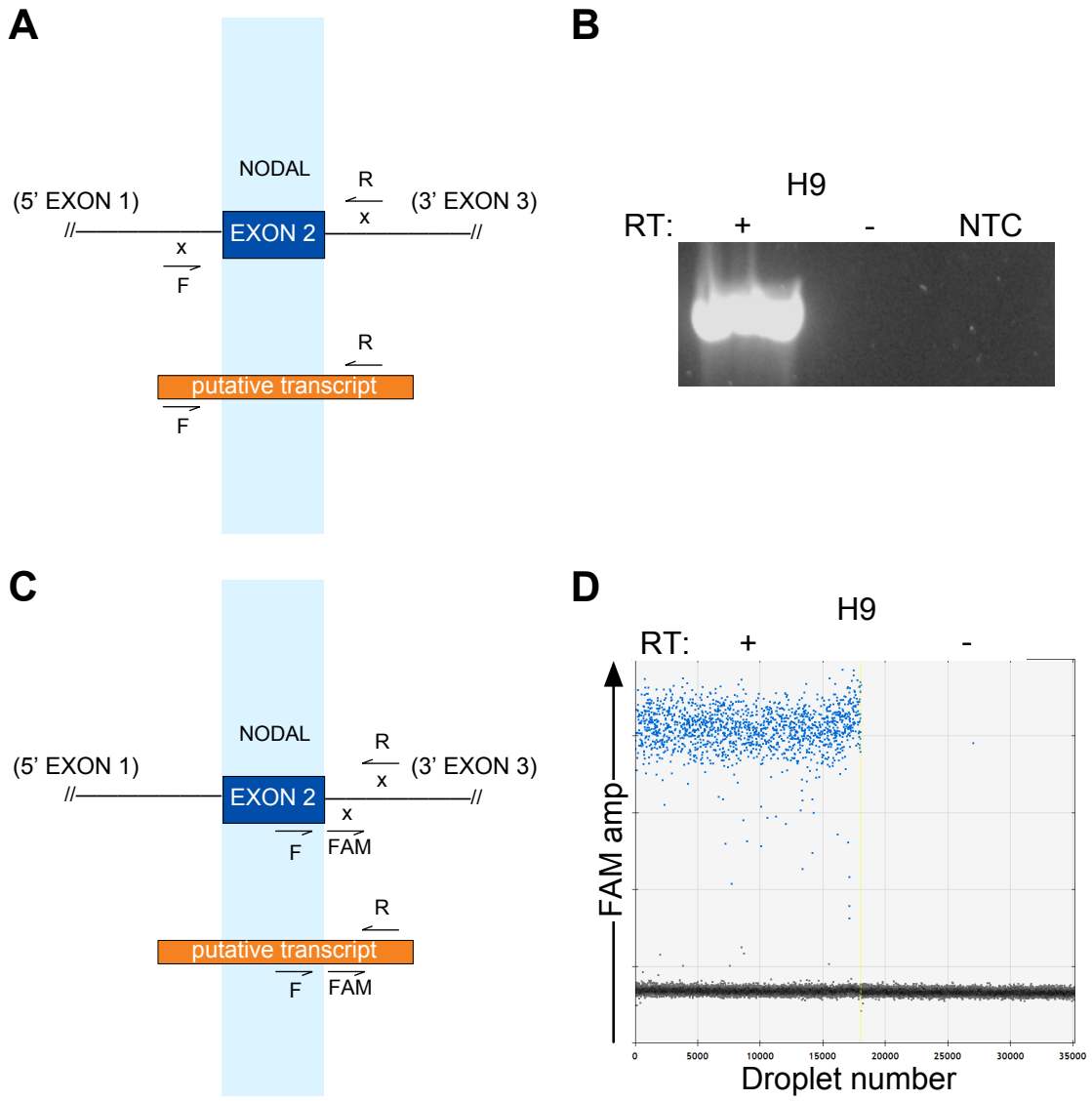


Figure 3.11: A transcript transcribed from a region encompassing the second constitutive exon of human *NODAL* is expressed in H9 hES cells. A) and C) Locations of primers used to amplify the putative transcript spanning constitutive exon 2 of human *NODAL*. The relative locations of *NODAL* exon 2 and the putative transcript are shown, with the light blue box indicating shared sequence. "F" = forward primer. "R" = reverse primer. "x" indicates no primer binding site in full-length *NODAL* transcripts. "FAM" = fluorescent probe. B) The putative transcript is detected in H9 hES cells. "NTC" = no template control. D) The ddPCR assay in (C) detects the putative transcript in H9 hES cells. "RT" = reverse transcriptase.

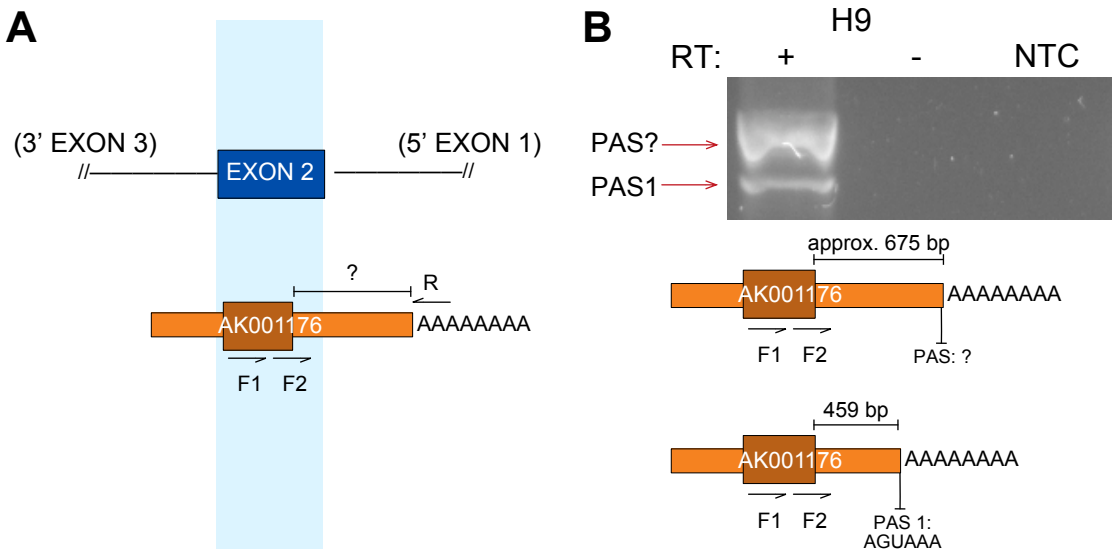


Figure 3.12: 3' RACE analysis of the putative transcript confirms antisense transcription relative to full-length *NODAL*, as well as alternative polyadenylation.

A) Primers used for 3' RACE analysis of the AK001176 transcript. Note that the orientation is flipped relative to figure 3.11, such that the putative open reading frame of AK001176 is shown left to right. The darker orange and thicker region of AK001176 indicates the putative open reading frame. The light blue box indicates shared sequence between *NODAL* exon 2 and the AK001176 transcript. "F1" = forward primer used for first round of PCR. "F2" = forward primer used for nested PCR. "R" = reverse adapter primer used for reverse transcription. B) AK001176 is alternatively polyadenylated. A nearby upstream common PAS for the longer PCR product could not be identified. The shorter PCR product resulted from polyadenylation at a more proximal AGUAAA PAS. "PAS" = polyadenylation site. "RT" = reverse transcriptase.

polyadenylated transcripts do not contain a known PAS [26]. This finding suggests that the AK001176 transcript is alternatively polyadenylated and confirmed that this transcript is transcribed from the opposite strand to *NODAL* and can thus be classified as an overlapping natural antisense transcript (NAT).

In the H9 "high" *NODAL* sample, signal from this assay was 23-fold less than that from the *NODAL* exon 2 assay (Figure 3.13A), suggesting that full-length *NODAL* was more highly expressed in this sample. In contrast, in the H9 "low" *NODAL* sample, signal from this assay was very similar (within two-fold) to that from exon 2 (Figure 3.13B), suggesting that the overlapping transcript contributed to the corresponding higher signal from exon 2 relative to other assays in this sample. I next compared expression of *NODAL* and the AK001176 transcript in several breast cancer cell lines. All of MCF7, T47D, and SUM 149 cell lines showed relatively high levels of *NODAL* according to the exon 2 assay, and extremely low levels of *NODAL* according to both exon 1 - exon 2 and exon 2 - exon 3 assays. The AK001176 transcript was detected at much higher levels than *NODAL* assays to exon 1 - exon 2 and exon 2 - exon 3, but at comparable levels to the assay for exon 2. Collectively, these results suggest that AK001176 NAT expression confounds analysis of full-length *NODAL* transcript within constitutive exon 2.

Since the signal from the AK001176 transcript was lower than that from the exon 2 assay in all samples tested, I was also interested in testing for expression of other transcripts containing exon 2 sequence. I discovered that *NODAL* exon 2 was an excellent candidate to form a circular RNA. Circular RNA forms when the 5' splice donor site of an intron forms a "back splice" with an upstream 3' splice site of the same or other exons in the transcript [11]. Relative to splice sites in general, it has been shown that circular RNA splice sites are more likely to be flanked by upstream and downstream intronic Alu repeat elements and that these genomic elements are more likely to be in opposite orientations. Single circularized exons were also found to be among the longest of all human exons, with an average length of 690 nucleotides [11]. In addition to constitutive exon 2 of *NODAL* being an extremely long exon (698 nucleotides), analysis of Alu repeats in the intronic sequences flanking *NODAL* exon 2 revealed two upstream Alu repeats and two downstream Alu repeats within 2 kb of *NODAL* exon 2 splice sites (Figure 3.14A).

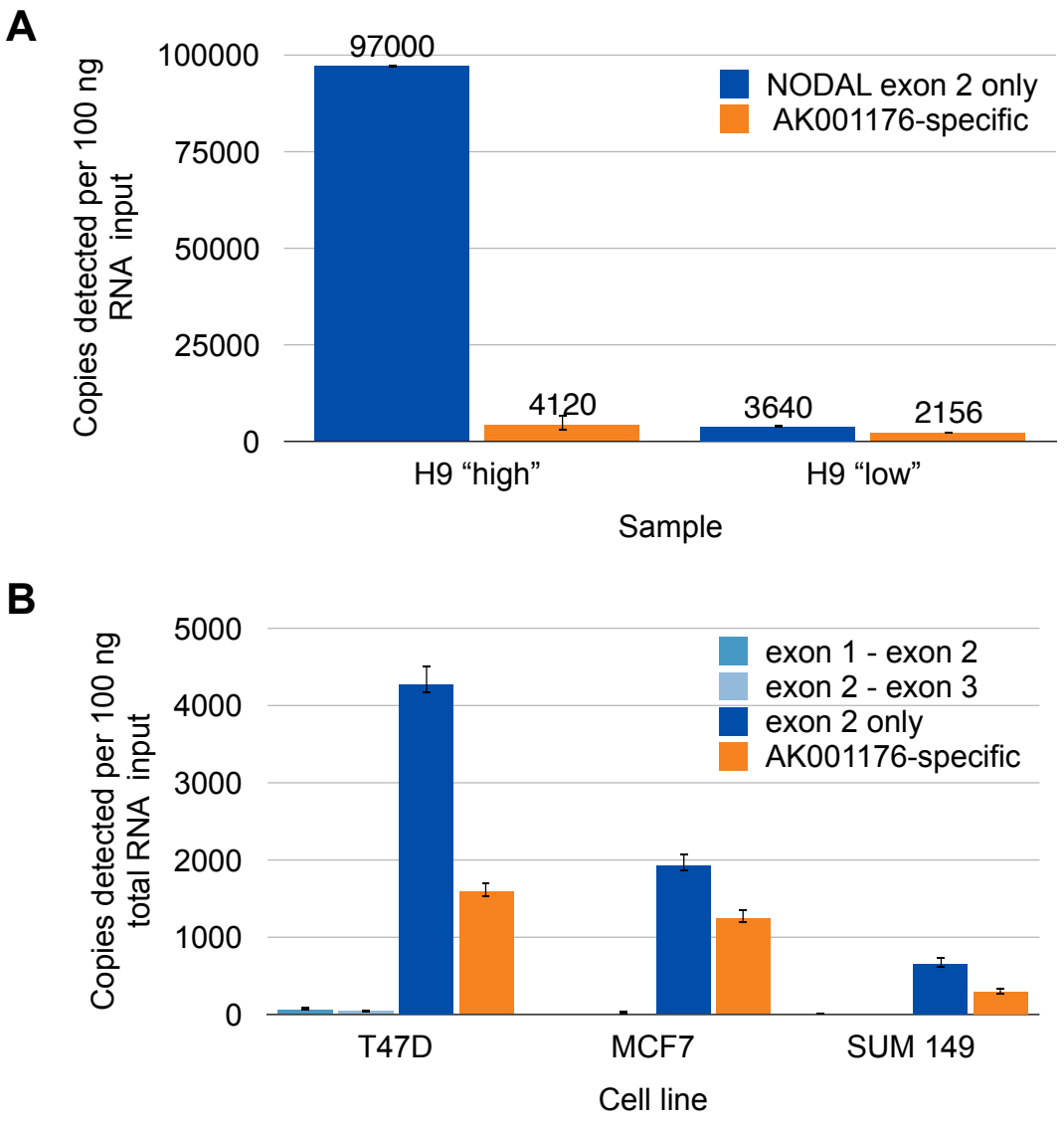


Figure 3.13: Expression of the *NODAL* natural antisense transcript (NAT) AK001176 in breast cancer cell lines.

A) Very different transcript levels were detected by the *NODAL* exon 2 assay and a NAT-specific assay in a sample with high *NODAL* transcript levels. Similar transcript levels were detected by these two assays in a sample with low *NODAL* transcript levels. B) Assays for both exon 2 and the NAT transcript detected high and similar transcript levels relative to exon boundary-spanning *NODAL* assays.

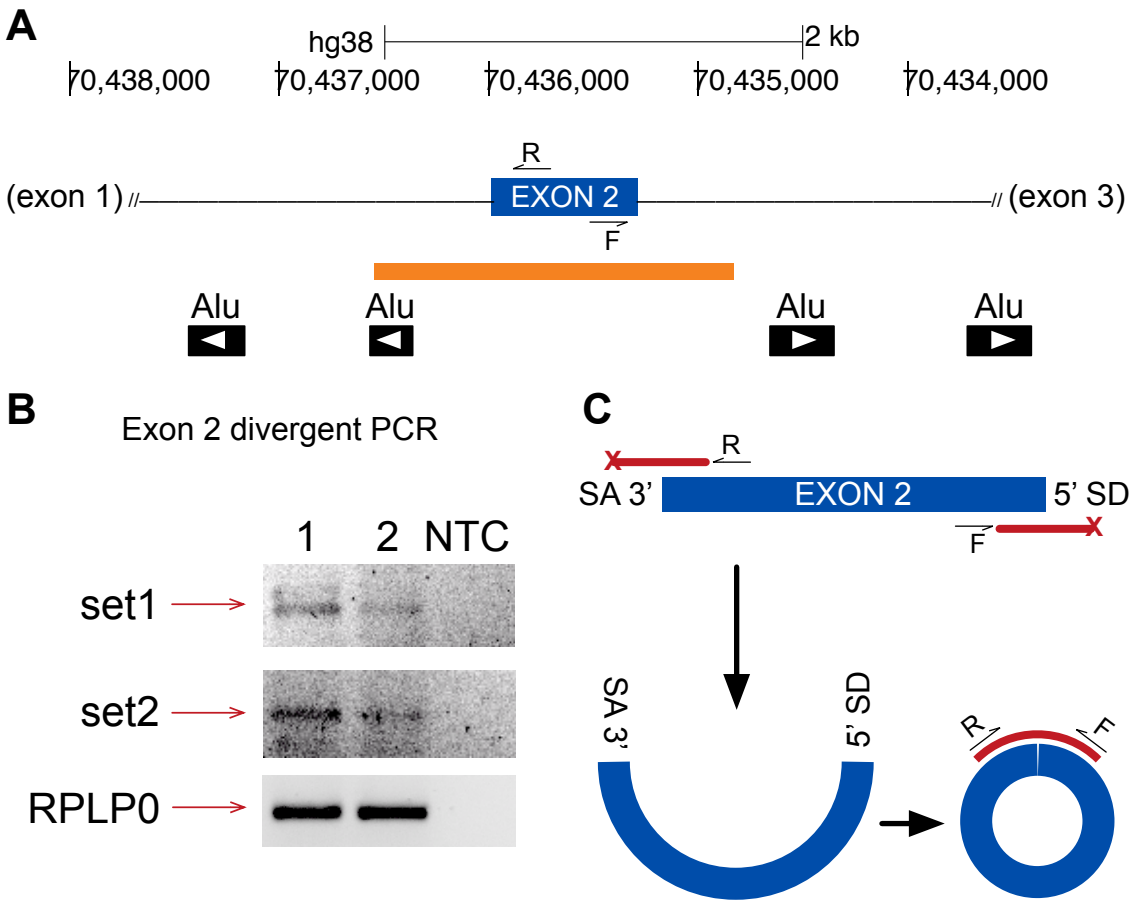


Figure 3.14: A circular RNA formed by the second constitutive exon of human *NODAL* is expressed in H9 hES cells.
A) Locations of Alu SINE elements from the repeat masker track of the UCSC human genome browser are shown relative to locations of *NODAL* exon 2 (blue) and the NAT transcript (orange). Locations of forward ("F") and reverse ("R") divergent primers are shown. Hg38 chromosome 10 coordinates and scale are shown at the top of the image. Arrows indicate orientation/strand of Alu elements. B) End-point PCR detection of circular exon 2 amplicons (and products resulting from template switching) with two different primer sets in two different H9 hES samples ("1" and "2"). "NTC" = no template control. Images were inverted for better visualization of bands. C) Schematic of *NODAL* exon 2 circular RNA and corresponding PCR strategy used. A back-splice of exon2 SD with the upstream exon 2 SA results in circular RNA formation. Red bars indicate PCR amplicons. "x" indicates non-productive amplification of linearly-spliced exon 2. "SA" = splice acceptor. "SD" = splice donor.

Moreover, Alu repeats in each intron had the same orientation, and were opposite in orientation relative to repeats in the adjacent intron. Divergent PCR of *NODAL* exon 2 revealed a single band for H9 hES cell RNA. Cloning and sequencing of this band confirmed the expression of a single exon circular RNA for *NODAL* exon 2 (Figure 3.14).

Having extensively characterized several transcripts within the *NODAL* locus, I was interested in further investigating the dynamics of some of these isoforms. First, RNA stability analysis was conducted to compare the dynamics of the two full-length alternatively spliced *NODAL* transcripts. Actinomycin D was used to block *de novo* transcription in H9 hES cells. Relative to long half-life *ACTB* transcripts, levels of control transcripts *MYC* and *TBP*, previously identified as having short half-lives [11, 35], were both significantly reduced after six hours of treatment with actinomycin D. These transcripts displayed first-order reaction-like kinetics indicative of a constant decay rate (Figure 3.15), validating the experimental approach used. Constitutive *NODAL* transcript was estimated to have a half-life of 5.0 hours (Figure 3.16A), while the estimated half-life for *NODAL* variant transcript was 8.9 hours (Figure 3.16B). The best fit curve for *NODAL* variant decay was a much poorer fit than that for constitutive *NODAL*, and the difference between *NODAL* variant and constitutive *NODAL* transcript half-lives was not statistically significant according to an analysis of covariance (ANCOVA) test (Figure 3.16C). Interestingly, constitutive *NODAL* transcript did have a half-life that was 2.5-fold longer than *MYC* (Figure 3.16D), in contrast to a genome-wide study in mouse ES cells which found *NODAL* and *MYC* to have very similar and very short (1.1 and 1.0 hours, respectively) half-lives [35].

Finally, to assess the impact of the alternative splicing of *NODAL* on human embryonic stem cell biology, a morpholino antisense oligonucleotide (MO) strategy was used to sterically block alternative exon splicing at the 5’ splice donor site (Figure 3.17A). Cells with high MO uptake were purified using FACS sorting and analyzed 48 hours after treatment. Relative to a control MO, cells treated with the alternative exon MO revealed an average 4.8-fold decrease in *NODAL* variant transcript expression (Figure 3.17B).

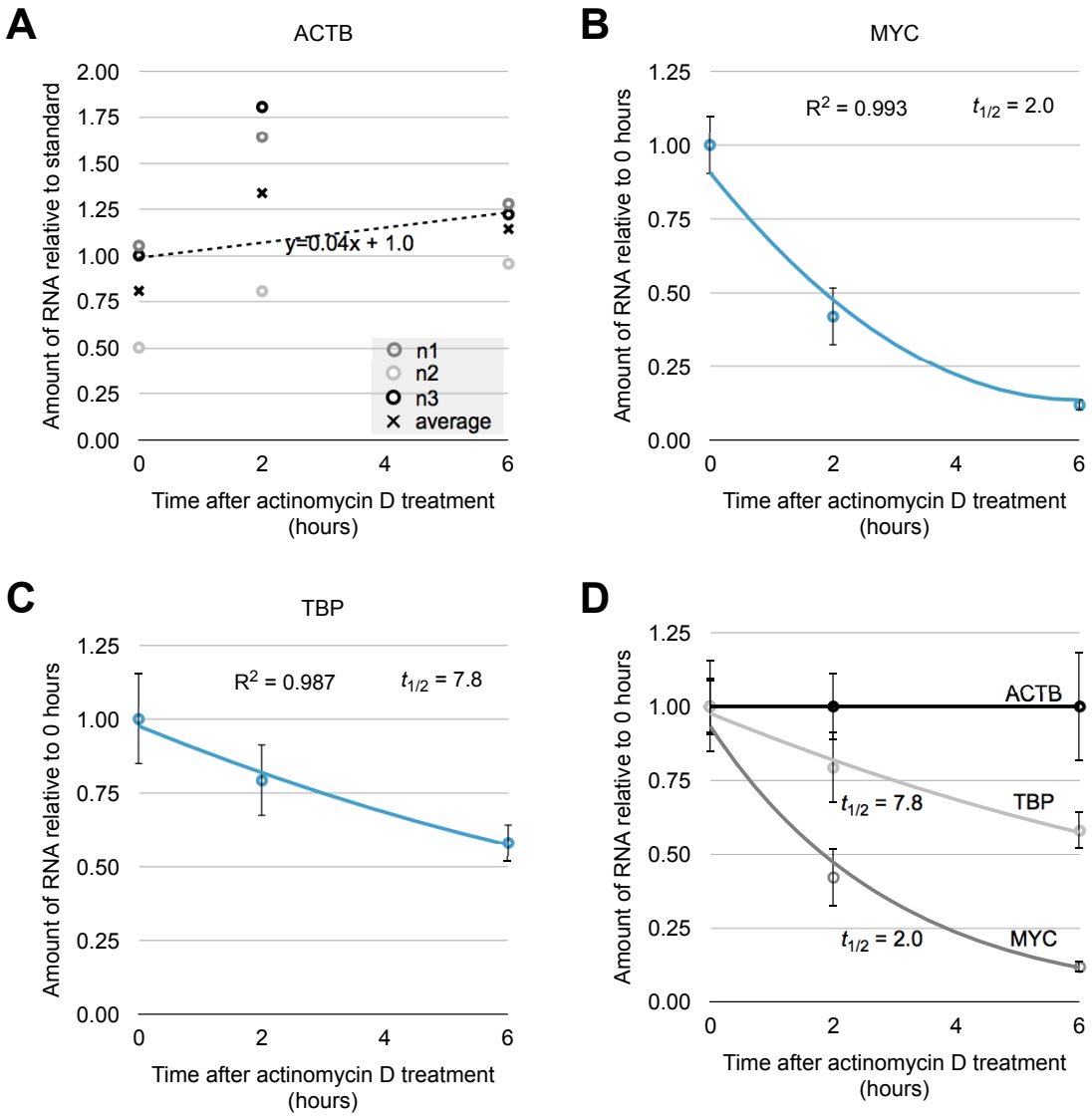


Figure 3.15: Actinomycin D treatment for assessing half-lives of RNA transcripts in H9 hES cells.

A) *ACTB* transcript levels are not decreased by Actinomycin D treatment over 6 hours and were used to control for differences in cell number between samples. Data points show transcript levels for three independent biological replications of the experiment (n1, n2, n3). "x" data points and the corresponding dashed linear regression line indicate average *ACTB* levels for each time point. B) Short half-life *MYC* transcript was used as a positive control to confirm RNA degradation. C) Degradation of *TBP* transcripts was also evident. D) Merged decay curves for positive controls *TBP* and *MYC* normalized to *ACTB*. Error bars indicate standard deviations. "$t_{1/2}$" = calculated half-lives in hours. "R^2" = coefficients of determination for exponential decay curves.

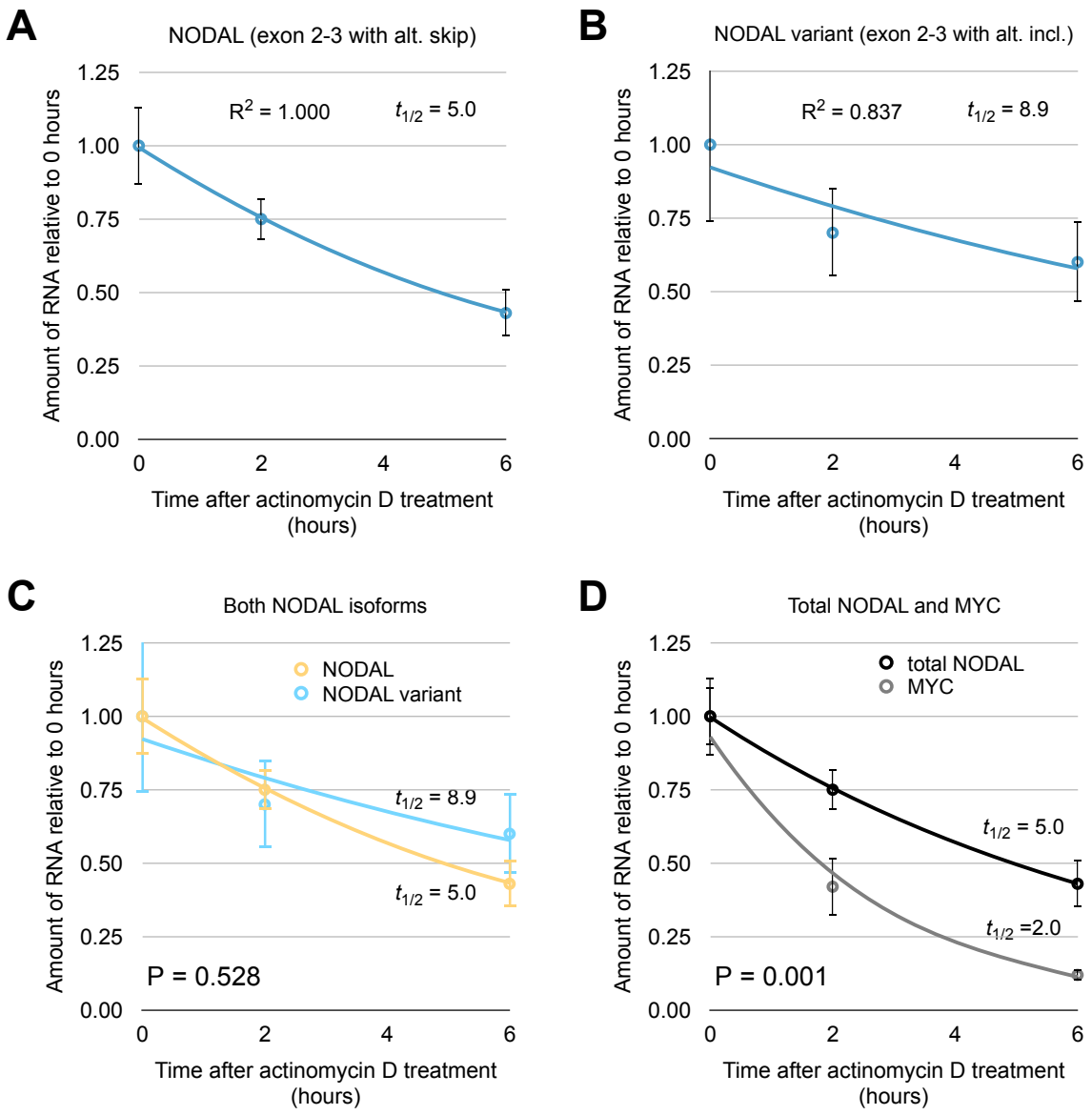
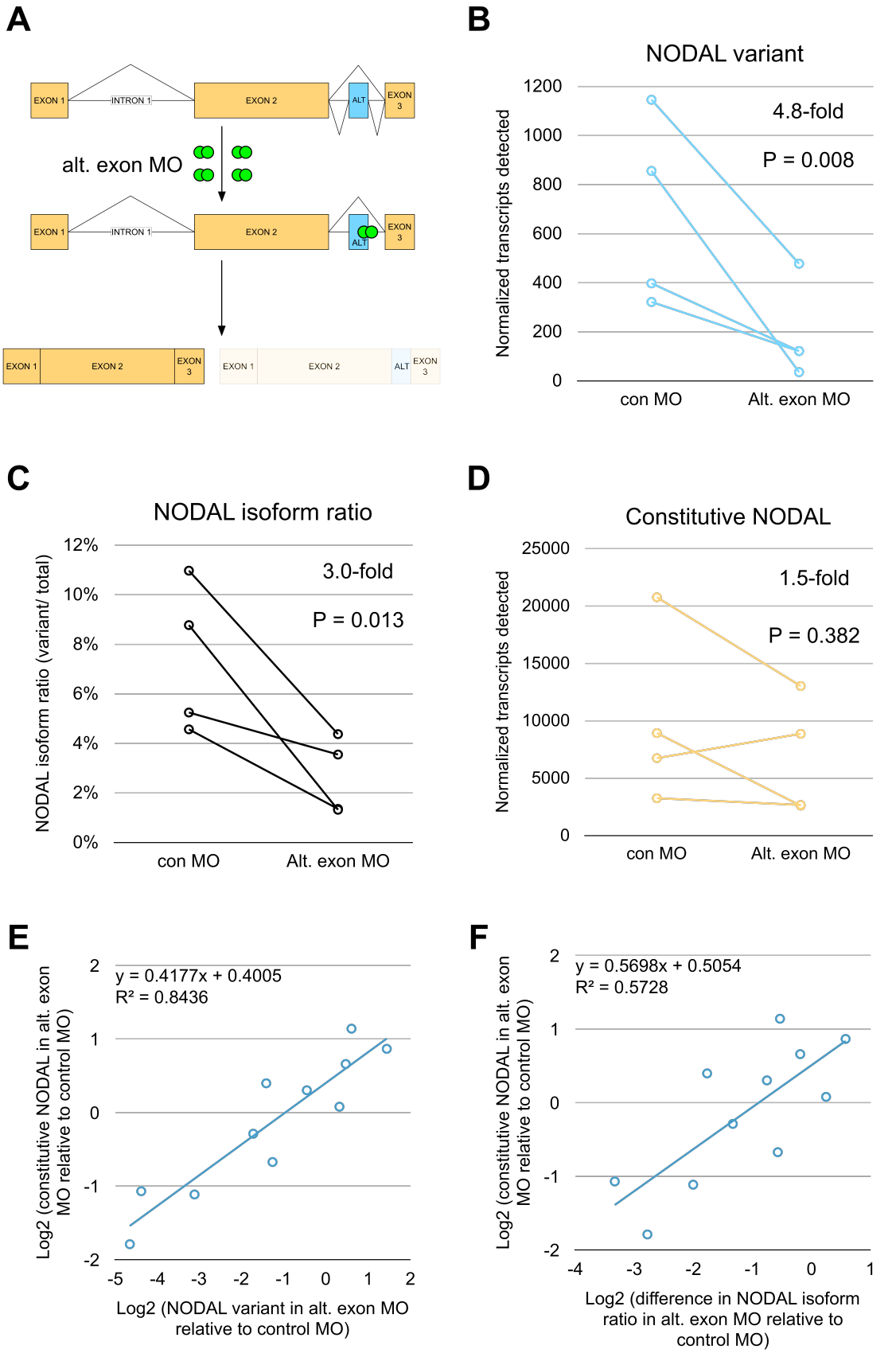


Figure 3.16: Both constitutive *NODAL* and *NODAL* variant transcripts are relatively stable in H9 hES cells.

A) Constitutive *NODAL* transcript decay revealed a half-life of 5.0 hours. B) *NODAL* variant transcript decay revealed a half-life of 8.9 hours. C) Comparison of decay for both alternatively spliced *NODAL* isoforms revealed similar half-lives. D) Comparison of decay for *NODAL* and *MYC* transcripts revealed substantially slower decay of *NODAL* transcripts. Error bars indicate standard deviations. "$t_{1/2}$" = calculated half-life. "R^2" = coefficients of determination for exponential decay curves. P-values show statistical significance results of analysis of covariance (ANCOVA) tests between transcripts.

Figure 3.17: *NODAL* variant knockdown and corresponding constitutive *NODAL* levels in H9 human embryonic stem cells (see over).
A) Schematic of experimental approach for *NODAL* variant-specific knockdown using a morpholino to block the alternative exon splice donor site and exon definition. Morpholino is represented by green circles. Splicing events are indicated by diagonal lines connecting exons. Spliced mRNA isoforms are shown at the bottom of the panel. B-C) Morpholino treatment was successful in consistently reducing levels of *NODAL* variant transcript. D) In three out of four replicates with FACS sorting, constitutive *NODAL* was reduced after *NODAL* variant knockdown. E-F) Analysis of 11 total experiments revealed reduced constitutive *NODAL* expression upon *NODAL* variant knockdown. Lines join control MO and alternative exon MO-treated samples from the same experiment. The geometric average decrease in transcript levels and P values indicating statistical significance results of paired t-tests for each transcript are shown for panels B-D. Coefficients of determination (R^2) and linear regression equations modelling the data are shown for correlation analyses in E-F. “alt.” = alternative. “con” = control. “MO” = morpholino.


A
EXON 1
INTRON 1
EXON 2
ALT
EXON 3
alt. exon MO
B
NODAL variant
Normalized transcripts detected
4.8-fold
P = 0.008
con MO
Alt. exon MO
C
NODAL isoform ratio
NODAL isoform ratio (variant/ total)
3.0-fold
P = 0.013
D
Constitutive NODAL
Normalized transcripts detected
1.5-fold
P = 0.382
E
y = 0.4177x + 0.4005
R² = 0.8436
Log2 (constitutive NODAL in alt. exon MO relative to control MO)
Log2 (NODAL variant in alt. exon MO relative to control MO)
F
y = 0.5698x + 0.5054
R² = 0.5728
Log2 (constitutive NODAL in alt. exon MO relative to control MO)
Log2 (difference in NODAL isoform ratio in alt. exon MO relative to control MO)

The ratio of *NODAL* variant to constitutive *NODAL* transcript also decreased an average of 3-fold, indicative of successfully altered alternative splicing of *NODAL* transcript (Figure 3.17C). There was no corresponding change in levels of constitutively spliced *NODAL* transcript according to a paired-t-test (P = 0.382; Figure 3.17D). To further assess a potential link between knockdown of the *NODAL* variant transcript and resulting constitutive *NODAL* expression levels, *NODAL* isoform levels were also measured in additional *NODAL* variant MO experiments that did not include FACS enrichment and displayed varying knockdown efficiencies. In a total of 11 separate experiments, there was a strong positive correlation between *NODAL* variant knockdown efficiency and reduced constitutive *NODAL* levels (Figure 3.17E). That is, the less *NODAL* variant there was after knockdown, the less corresponding constitute *NODAL* was present. This effect was not solely the result of less total *NODAL* to begin with in the samples with ostensibly "efficient" *NODAL* variant knockdown: There was also a strong positive correlation between *NODAL* variant knockdown efficiency, as measured by the relative *ratio* of *NODAL* variant to constitutive *NODAL*, and the extent to which corresponding constitutive *NODAL* levels were reduced (Figure 3.17F).

To compare *NODAL* variant knockdown to knockdown of total *NODAL*, a second MO targeting the 5' splice donor site of constitutive exon 2 was designed (Figure 3.18A). Identical parallel treatment with this MO resulted in an average 4.2-fold reduction in constitutive *NODAL* transcript levels relative to control MO-treated cells (Figure 3.18B). As expected, *NODAL* variant transcript levels were also reduced by an average of 4.4-fold (Figure 3.18C). However, the *NODAL* isoform ratio was unchanged, indicative of uniform knockdown of total *NODAL* transcript (Figure 3.18D).

To assess the potential broad impact of *NODAL* variant expression on hES cell biology, I first selected the MO experiment with the most efficient *NODAL* variant knockdown for expression analysis of genes involved in human embryonic stem cell self-renewal and differentiation using a PCR array. Relative to control MO-treated cells and using a 2-fold change as a cutoff for differential gene expression, a 96% or 25-fold reduction in

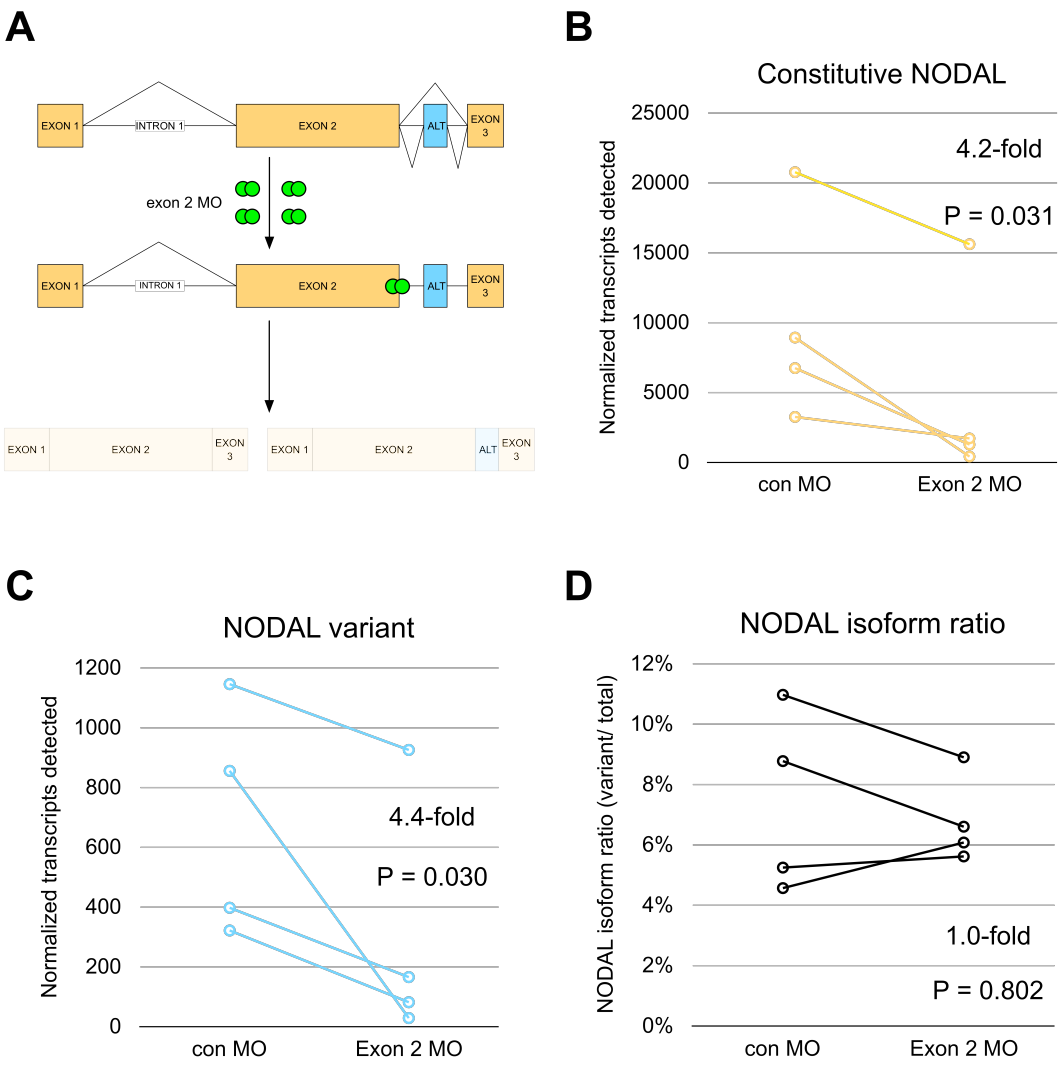


Figure 3.18: Total *NODAL* knockdown in H9 human embryonic stem cells. A) Schematic of experimental approach for total *NODAL* knockdown using a morpholino to block the constitutive exon 2 splice donor site and exon definition. Morpholino is represented by green circles. Splicing events are indicated by diagonal lines connecting exons. Spliced mRNA isoforms are shown at the bottom of the panel. Morpholino treatment was successful in consistently reducing levels of constitutive *NODAL* (B) and *NODAL* variant (C) transcript. D) The *NODAL* isoform ratio (*NODAL* variant/ total *NODAL*) remained unchanged upon morpholino treatment. Results are shown for four independent biological replications of the experiment. Lines join control MO and exon 2 MO-treated samples from the same experiment. The geometric average decrease in transcript levels is indicated in the top right corner of each panel. Below this are P values indicating statistical significance results of paired t-tests for each transcript.

NODAL variant transcript resulted in altered expression for 15% of genes tested (Figure 3.19A,B). The same analysis in another replicate of the experiment where 62% or 3-fold reduction in *NODAL* variant transcript levels were achieved yielded differential expression for 8% of genes tested. While there was little overlap in the identity of the genes with altered expression across experiments (Figure 3.19B), genes with altered expression upon *NODAL* variant MO treatment tended to also be similarly altered upon parallel treatment with exon 2 MO where total *NODAL* levels were comparably reduced (Figure 3.19A-C). This effect was consistent across both independent experiments.

Knockdown of total *NODAL* transcript induced more widespread changes in expression of genes related to embryonic stem cell identity (Figure 3.20). Notably, several genes involved in the maintenance of pluripotency and embryonic stem cell identity were downregulated in response to total *NODAL* knockdown. These included master regulators of ES cell pluripotency such as *TERT* [36] and *MYC* [37], and genes involved in NODAL signalling such as *GDF3*, *TDGF1* (Cripto) and *SMAD3* (Figure 3.20B). These results are consistent with a role for *NODAL* in the maintenance of ES cell pluripotency. Relative to knockdown of the *NODAL* variant, knockdown of total *NODAL* induced more changes in gene expression (Figure 3.20C). A reduction of total *NODAL* by 96% or 23-fold resulted in altered expression for 28% of genes tested, while a reduction of total *NODAL* by 82% or 5-fold resulted in altered expression for 20% of genes tested.

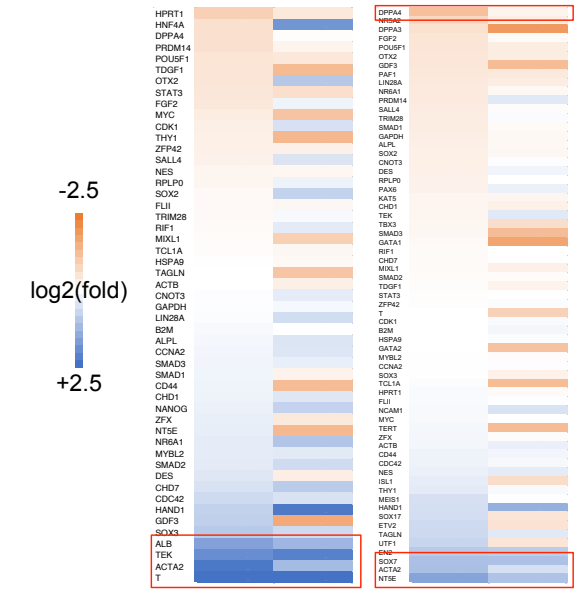
3.3 Discussion

This chapter detailed the characterization of specific transcript isoforms for the human *NODAL* locus. In addition to further study of the newly identified genetically-regulated splice variant reported in the previous chapter, I reported detection of an alternative transcriptional start site and putative first exon upstream of constitutive exon 1, alternative polyadenylation site usage, a circular RNA from constitutive exon 2, and confirmed expression of an antisense transcript encompassing the constitutive exon 2 locus. Collectively, these results point to complex regulation of *NODAL* gene expression at the RNA level that can now be used to guide more precise and accurate assay

Figure 3.19: Effect of *NODAL* variant-specific knockdown in H9 hES cells on expression of genes related to pluripotency and differentiation (see over).

A) Heat maps for differences in gene expression between *NODAL* MO treated and control treated hES cells. Data from two experiments with the most efficient *NODAL* variant knockdowns are shown. Differences are ranked from most decreased (top; red) to most increased (bottom; blue) after MO treatment for the *NODAL* variant splice blocking MO. Corresponding changes in *NODAL* exon 2 MO treated cells are shown to the right. Red boxes indicate genes with log2(fold-change > 1), corresponding to a fold-change > 2. “rep” = biological replicate. “MO” = morpholino used. “K.D.%” = percentage knockdown (e.g. 100% = no detectable transcript, 0% = no knockdown relative to control). “K.D. fold” = knockdown as fold-change relative to control levels (e.g. 2-fold = 50% knockdown). B) Genes with > 2-fold difference between control MO and *NODAL* alternative exon MO. Note: genes with non-specific PCR results in any samples were not included in this figure. C) Correlation of gene expression responses to *NODAL* variant knockdown and total *NODAL* knockdown for genes in (B). Lines represent linear regression equations modelling the data. Coefficients of determination (R^2) are shown.

A



B

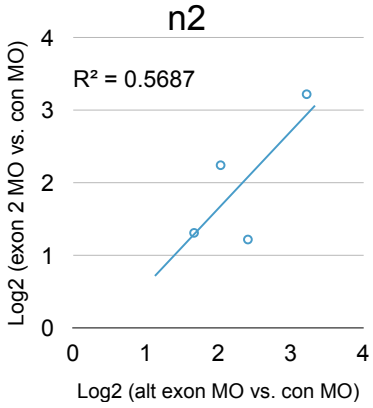
n2 > 2-fold

Gene	Log2 fold-change Alt exon MO vs. con MO	Log2 fold-change Exon 2 MO vs. con MO
ALB	1.67	1.31
TEK	2.03	2.25
ACTA2	2.41	1.22
T	3.22	3.22

n4 > 2-fold

Gene	Log2 fold-change Alt exon MO vs. con MO	Log2 fold-change Exon 2 MO vs. con MO
DPPA4	-1.20	-0.20
SOX7	1.12	1.10
ACTA2	1.15	0.55
NT5E	1.63	1.07

C



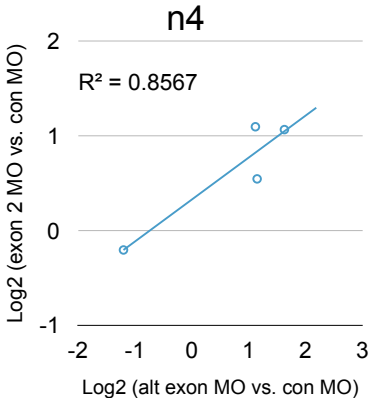
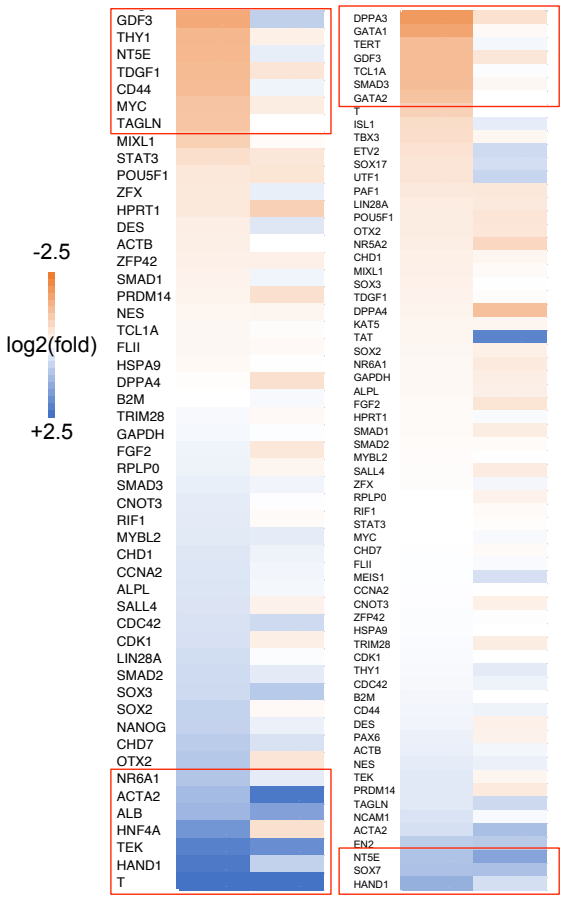


Figure 3.20: Effect of total *NODAL* knockdown in H9 hES cells on expression of genes related to pluripotency and differentiation (see over). A) Heat maps for differences in gene expression between *NODAL* MO treated and control treated hES cells. Data from two experiments with the most efficient total *NODAL* knockdowns are shown. Differences are ranked from most decreased (top; red) to most increased (bottom; blue) after MO treatment for the *NODAL* exon 2 splice blocking MO. Corresponding changes in *NODAL* alternative exon MO treated cells are shown to the right. Red boxes indicate genes with log2(fold-change > 1), corresponding to a fold-change > 2. “rep” = biological replicate. “MO” = morpholino used. “K.D.%” = percentage knockdown (e.g. 100% = no detectable transcript, 0% = no knockdown relative to control). “K.D. fold” = knockdown as fold-change relative to control levels (e.g. 4-fold = 75% knockdown). B) Genes with > 2-fold difference between control MO and *NODAL* exon 2 MO. C) The proportion of genes with substantially altered expression resulting from *NODAL* MO treatment is greater for *NODAL* exon 2 MO treated cells relative to *NODAL* alternative exon treated cells in both replicates.

A



B

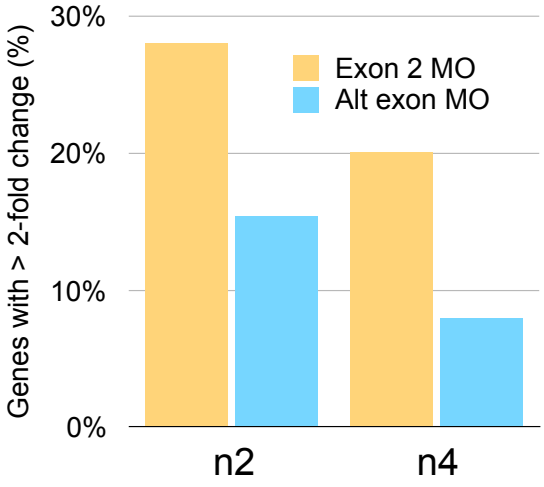
n2

Gene	Log2 fold-change exon2 MO vs. con MO
GDF3	-1.63
THY1	-1.38
NT5E	-1.35
TDGF1	-1.30
CD44	-1.28
MYC	-1.12
TAGLN	-1.10
NR6A1	1.03
ACTA2	1.22
ALB	1.31
HNF4A	1.89
TEK	2.25
HAND1	2.38
T	3.22

n4

Gene	Log2 fold-change exon 2 MO vs. con MO
DPPA3	-1.94
GATA1	-1.73
TERT	-1.28
GDF3	-1.27
TCL1A	-1.27
SMAD3	-1.26
SCN1A	-1.22
GATA2	-1.14
NT5E	1.07
SOX7	1.10
HAND1	1.41

C



utilization and enrich our modelling of *NODAL* biology in human pluripotent stem cell and cancer cell lines.

I found the alternatively spliced *NODAL* exon identified in the previous chapter to contribute to full-length *NODAL* transcripts containing an ORF with a distinct C-terminus relative to constitutive NODAL. This variant transcript is fully spliced and polyadenylated. Mammalian transcripts are targeted for nonsense mediated decay (NMD) if a premature termination codon (PTC) is present more than 50 bases upstream of an exon-exon junction complex according to the "EJC" model, or possibly if they result in very long 3' UTRs according to the "faux 3' UTR" model [38-40]. Notably, a recent genome-wide survey of NMD found widespread evidence of the EJC model explaining instances of NMD, but did not find consistent evidence of the faux 3' UTR model influencing NMD in human cells [38]. Inclusion of the alternative *NODAL* exon alters the downstream translational reading frame relative to the constitutively spliced isoform, resulting in a premature termination codon (PTC) just downstream of the 5' end of constitutive exon 3. However, the *NODAL* variant PTC is in the most 3' and final exon and thus not upstream of any exon-exon junction, and is less than 150 bases upstream of the constitutive *NODAL* stop. Thus, the *NODAL* variant transcript is not a good candidate to induce NMD. Results of an RNA stability experiment where *NODAL* variant transcripts were as stable, and possibly more stable, than constitutive *NODAL* transcripts indicated that *NODAL* variant transcripts are not targeted for rapid degradation. Collectively, these results suggest the *NODAL* variant is processed and likely translated in a similar fashion to annotated *NODAL*. However, whether *NODAL* variant splicing induces a NMD response was not directly assessed. Determination of the exact transcriptional start site(s) used by *NODAL* transcript isoforms was confounded by potential incomplete reverse transcription of cDNA. A 5' cap-specific method such as RLM RACE will be used in future studies for exact 5' end determination. The enrichment of an alternative first exon and skewed polyadenylation site usage for the *NODAL* variant relative to constitutive *NODAL* suggests there is coordinated regulation of the *NODAL* variant transcript that extends beyond the direct definition of a splice donor site formed by the rs2231947 T allele. This apparent coordination with polyadenylation is interesting given that a link between alternative polyadenylation and

alternative splicing has been described, but only for 3' terminal exon selection [18]. To my knowledge, processes that link alternative splicing events at non-adjacent exons and to alternative transcriptional start sites have not been described before in the literature.

The ability to specifically analyze *NODAL* variant and constitutively spliced *NODAL* transcripts on a single molecule basis was potentiated by a duplexed ddPCR assay that offers several benefits over traditionally employed splice variant detection methods such as those used in the previous chapter. In the case of exon skipping events, the most commonly employed detection method utilizes one set of primers targeting constitutive exons flanking the alternative splicing event of interest. This results in the amplification of a shorter amplicon corresponding to the alternative exon-skipping isoform, and a longer amplicon corresponding to the alternative exon including isoform. Due to the difference in amplicon length, these two products can be easily resolved using agarose gel electrophoresis. This is a major advantage of this method, as it provides high confidence in the specific detection of each isoform. However, this method relies on densitometry-based analysis of end-point PCR products. This is not ideal for quantitative purposes since the quantity of an amplicon at the end of a PCR reaction is not necessarily indicative of the initial relative target quantity in the sample due to reaction plateauing and potentially different amplification efficiencies between amplicons that skew the final relative abundance. Signal detection is also prone to saturation, and for a molar comparison of the splice variant ratio, the signal needs to be corrected based on relative differences in the size of the amplicons. Lastly, isoforms with low expression may not be easily detected. The amplification efficiencies and thus resulting splice ratios are also potentially influenced by reaction conditions such as choice of annealing temperature and salt concentration in the PCR reaction [41].

Real-time PCR offers a suitable quantitative method. However, such assays are most reliable when implemented as separate assays for detection of each splice variant, and require the inclusion of standard curves of known quantity for each variant to be detected, so that accurate relative comparisons can be made. Thus, while more quantitative, these assays are more labour-intensive and their reliance on standard curves presents more opportunity for technical sources of error in splice variant quantification. Additionally, it

is difficult to be confident in the specificity of an assay targeting an exon skipping event, as primers designed to splice junctions may also effectively amplify transcripts with alternative exons in some contexts [41].

The advent of digital droplet PCR (ddPCR) offers an alternative strategy for splice variant detection that is easily duplexed as in end-point detection methods, but also quantitative as in real-time methods. Moreover, ddPCR methods are absolutely quantitative, not influenced by amplification efficiency, and do not require standard curves. Furthermore, ddPCR has single molecule resolution, and duplexing in a splice variant assay allows confident isoform classification on an individual transcript basis. The results presented here illustrate the power of a *NODAL* splicing ddPCR assay to quantify constitutive *NODAL* and *NODAL* variant transcripts to deliver accurate isoform-specific quantification in knockdown and RNA stability assays.

NODAL signalling is generally thought to be essential for human embryonic stem cell pluripotency [42, 43] and consistently high levels of *NODAL* expression have been reported for this cell type [44]. Here I have reported extremely low *NODAL* expression at the RNA level in some isolates of hES cells with typical pluripotent stem cell morphology and expression of markers of pluripotency, cultured continuously under standard feeder-free defined culture conditions. It is possible that low *NODAL* levels are indicative of cultures poised for (or already undergoing) early differentiation. In this vein, *NODAL*, *LEFTY1*, and *LEFTY2* displayed some of the most rapid down-regulation upon spontaneous hES cell differentiation in a small panel of pluripotency markers [45]. Still, it is also possible that high *NODAL* expression is not strictly required for the maintenance of pluripotency, and that there are redundant or compensatory mechanisms that can sustain pluripotency in the absence of *NODAL*. Studies have shown that exogenous *NODAL* reduced or delayed spontaneous differentiation of human embryonic stem cells, but the effects of directly knocking down *NODAL* in hESCs have never been reported. In addition, the experimental modelling used in many studies of *NODAL* in hES cell biology is not specific to *NODAL*. For example, the ALK4/5/7 inhibitor SB431542 [46] is often used to study hES cell fate (e.g. [42, 43, 47]), but results in broad inhibition of NODAL, Activin, TGF-beta, and other superfamily member signalling through any of these

receptors. When NODAL was more specifically inhibited by Vallier and colleagues [43] by over-expressing or treating hES cells with Lefty or Cerberus-short, the latter of which has been identified as a NODAL-specific inhibitor in xenopus [48], widespread differentiation was not observed. TRA-160-positive cells were still evident (although reduced in frequency for Lefty-over-expressing cells), and stable hES cell colonies were as efficiently generated from cells over expressing Lefty and Cerberus-short relative to GFP-expressing control cells. This, together with the findings presented here suggest that there may be redundancy in the signaling pathways required for the maintenance of pluripotency. Indeed, TGF-beta1 and Activin have been identified as likely candidates to complement NODAL signalling. Both genes were found to be highly expressed in both MEF feeder cells and hES cells, while *NODAL* was easily detected in hES cells but not MEFs.

More generally, this study supports the notion that a pluripotent gene expression signature is not static or universal, but rather partially stochastic, and that the combinations of active transcription factor networks and signalling pathways that can support the pluripotent state can drift with culture conditions and microenvironmental factors, between cell lines, and due to other unknown variables. Indeed, it has been suggested that there exists a spectrum or continuum of pluripotent states both *in vitro* and *in vivo* (reviewed by [49]). This property of ES cells may also explain why efficient knockdown of *NODAL* did not result in especially robust and consistent changes in specific genes between replicate experiments conducted on cells from subsequent passages. In this fashion, NODAL may maintain different gene expression networks dependent on the cellular context. Since the knockdown efficiency was different between experimental replicates, dosing and threshold effects could also result in different impacts on gene expression after *NODAL* knockdown.

The observed variability in hES cell *NODAL* transcript levels was certainly staggering. That *NODAL* transcript levels were dramatically and reversibly influenced by culture conditions may be an indication that more general differences exist between hES cells cultured in MEF-conditioned serum replacement-based media and in more defined media. In general, some differences in hES cells cultured under varying conditions have

been observed [50, 51]. However, very little work has directly compared culture with MEF-conditioned media and defined media such as mTESR [52]. Perhaps surprisingly, there is a striking absence of work involving comprehensive and quantitative profiling of hES gene expression signatures between MEF-conditioned media and defined media, to investigate to what extent such media may promote distinct pluripotent states. While the choice of culture media did have a dramatic effect on *NODAL* expression, the magnitude of this effect still paled in comparison to the overall variability observed in *NODAL* expression, suggesting it may not be the only or even major factor dictating *NODAL* expression in hES cells.

In addition to the unexpectedly low levels of *NODAL* transcript sometimes observed in hES cells, I also made the somewhat surprising discovery of especially low *NODAL* expression at the transcript level across numerous human cancer cell lines. The extremely low or sometimes virtually undetectable levels of *NODAL* transcript reported here for cell lines such as C8161 aggressive melanoma and MDA-MB-231 triple-negative breast cancer are inconsistent with functional studies in these cell lines where *NODAL* knockdown, mediated at the RNA level through RNA interference [31, 32], or inhibition of endogenously expressed protein [53] resulted in profound phenotypic effects. Notably, *NODAL* mRNA expression was not reported in these papers, so it is difficult to tell whether the detectable and functionally relevant levels of NODAL protein reported in these studies were expressed from cells with considerably higher *NODAL* mRNA levels, or if NODAL protein is perhaps generally translated or stabilized at high levels despite universally low mRNA levels. Another possibility is that there is only a minority population of cells expressing *NODAL*. At least one other group has reported undetectable *NODAL* transcript expression in MDA-MB-231 cells when using a real time PCR assay spanning an exon-exon boundary [54].

Estimates for the quantity of total RNA per cell (1-50 pg/cell) suggest that 100 ng total RNA represents about 4,000 cells. Thus, samples for which 1-2 copies of *NODAL* transcript are detected is indicative of expression of a single transcript for every several thousand cells. Even if inefficiencies in RNA extraction and reverse transcription are considered, this undoubtedly represents extremely low gene expression, or is indicative

of only a subpopulation of cells that are producing *NODAL*. It is not likely that technical inefficiencies can fully account for these results, as hES samples with high *NODAL* expression have been repeatedly analyzed in parallel. Multiple priming strategies for reverse transcription and different PCR detection assays also consistency revealed low expression in cancer cell line samples. Lastly, RNA integrity analysis shows intact RNA with no signs of degradation, and high levels of house-keeping genes such as *RPLP0* and other genes of interest have been routinely amplified from all of the low *NODAL*-expressing samples presented here. *NODAL* levels in different RNA samples from the same cancer cell line were also variable. For example, other isolations of RNA from the T47D breast cancer cell line have been found to have as few as 1-2 copies of total *NODAL* transcript per 100 µg RNA, while the SKOV3 ovarian carcinoma cell line which had no expression in the sample shown here has revealed more than 10 copies of total *NODAL* transcript in 100 ng RNA from other isolates.

One other study has directly compared *NODAL* expression levels between two cancer cell lines and hES cells with multiple *NODAL* assays, although this analysis was conducted using semi-quantitative end-point PCR [24]. For the C8161 cell line, an assay crossing the exon 2 - exon 3 boundary resulted in a very low intensity band that was barely detectable. In contrast, an assay internal to exon 2 yielded a band of much higher intensity. This result is consistent with those presented here, which reveal that this higher expression may at least partially result from expression of an antisense transcript sharing sequence with exon 2 and solely the constitutive *NODAL* transcript. The increased signal from assays internal to exon 2 is likely not the result of higher reverse transcription efficiency in this region of the transcript since signal was fairly uniform across all regions of the transcript in an H9 sample with high *NODAL* expression. It is also possible that unspliced pre-mRNA is a source of higher signal within constitutive exon 2. Although this possibility was not assessed here, the presence of such RNA has been shown for transcripts of *NODAL*-related genes in zebrafish [55]. Indeed, unless polyA tail-specific reverse transcription is performed, or PCR assays cross exon-exon boundaries, it is not possible to distinguish unspliced pre-mRNA from processed transcripts. Nonetheless, assays targeting only exon 2 of *NODAL* are not specific to full-length processed *NODAL* transcript. However, such assays have been widely employed when assessing *NODAL*

levels in many publications (e.g. [24, 31, 56-62]). Going forward, it is highly recommended that specific assays for the antisense transcript, circular RNA, potential unspliced pre-mRNA, and constitutive *NODAL* be employed to untangle the contributions of each transcript to any overall change in expression measured by the assays internal to exon 2. It will be interesting for future studies to explore whether these three transcripts show similar responses to altered microenvironments, and if over-expression of *NODAL* affects levels of antisense or circular RNAs, and vice-versa. Interestingly, the antisense transcript was found to be polyadenylated. This transcript also contains an ORF and is predicted to code for a protein with an N-terminal signal peptide, suggesting it is likely translated (Appendix A). However, since it shares coding sequence with much of exon 2, outside of *NODAL* codon wobble sites, much of its coding potential is likely influenced and constrained by the highly conserved constitutive *NODAL* coding sequence.

Interestingly, while *NODAL* was identified as having one of the shortest half-lives (just over 1 hour) in a global analysis of mRNA stability in mouse ES cells [35], our data revealed a half-life of over five hours in human ES cells. It is difficult to compare absolute half-lives between these studies, as different normalization strategies and other experimental variables can be confounding. However, the mouse ES study also identified the *MYC* transcript as having a very similar half-life to *NODAL* in their system, with a half-life of just under one hour. Inclusion of *MYC* in the analysis presented here revealed a similar half-life of two hours, substantially shorter than that of *NODAL*. A more comprehensive analysis of multiple transcripts would need to be conducted to determine if *NODAL* mRNA is more stable in human ES cells. If so, it is possible this is a result of altered mRNA stability in the primed, epiblast stem cell-like state of human ES cells relative to their naive “ground-state” mouse counterparts, although analysis of the raw data from [35] did not reveal substantially increased (>2-fold) stability of *NODAL* transcript when early differentiation was induced in mouse ES cells. Notably, *NODAL* variant transcript was found to be at least as stable as constitutive *NODAL* in hES cells. This finding suggests that lower levels of the *NODAL* variant relative to constitutive *NODAL* does not result from lower relative stability, and that the *NODAL* variant is not a

"mis-spliced" mRNA that is rapidly targeted for degradation by RNA surveillance pathways.

It is tempting to conclude that since constitutive *NODAL* levels correlated well with *NODAL* variant levels after specific knockdown of the latter that the *NODAL* variant promotes or maintains expression of *NODAL* in general. There are several possible factors that could contribute to the observed results. First, since *NODAL* variant is spliced as a proportion of total *NODAL* transcribed, samples with lower levels of *NODAL* will appear to have a better absolute knockdown partially due to stochastic variability of *NODAL* levels. This factor could not explain the entire effect, as experiments resulting in lower ratios of *NODAL* variant to constitutive *NODAL* also resulted in lower constitutive *NODAL* levels. Another variable is the effectiveness of the MO at blocking splicing, as it cannot be assumed that this is constant. If there is a causal relationship between alternative exon MO treatment and constitutive *NODAL* splicing or expression, it would be interesting to explore if this effect is dependent on alternative exon splicing per se, or if targeting the alternative exon splice donor site interferes with an element such as an intronic splicing enhancer for constitutive exon 2. This effect could also be the indirect result of *NODAL* variant expression involving a positive feedback on *NODAL* expression in general.

Genes with altered expression after *NODAL* variant knockdown tended to be altered in a similar fashion upon knockdown of total *NODAL* transcript. However, the number of genes with altered expression was much lower upon *NODAL* variant knockdown relative to total *NODAL* knockdown. It is possible that the *NODAL* variant shares limited functional redundancy with constitutive *NODAL*, or that the resulting decrease in constitutive *NODAL* levels is sufficient to induce a partial and similar response. More efficient knockdown of both total *NODAL* and *NODAL* variant in the "n2" experiment induced gene expression changes in a higher percentage of genes tested relative to an experiment with less efficient knockdowns for both total *NODAL* and *NODAL* variant. It is also unsurprising that total *NODAL* knockdown induced more changes in gene expression in both experiments than *NODAL* variant knockdown, as a genetically regulated splice variant sharing only partial sequence with constitutive *NODAL* is likely

to have a more limited impact on ES cell biology than a highly conserved bona fide regulator of stem cell fate. Despite achieving very efficient knockdown, these experiments were limited in their potential disruption of NODAL signalling: Only a portion of the total cell population was analyzed, and as a secreted paracrine growth factor, NODAL from cells receiving a low morpholino dose could still signal to the cells receiving a high morpholino dose that were analyzed. Furthermore, this experimental scheme is limited in its duration, and did not allow for sustained long-term reduction of *NODAL* levels.

In summary, this chapter identified several distinct transcripts expressed from the *NODAL* gene locus. At least two of these transcripts, the constitutive and alternatively spliced isoforms introduced in the previous chapter, exist as full-length spliced and polyadenylated stable transcripts containing open reading frames. The next chapter will characterize the translated products of these isoforms to examine how alternative splicing impacts NODAL biology at the protein level.

3.4 Methods

3.4.1 RNA extraction

Total RNA was isolated from cultured cells using the PerfectPure RNA Cultured Cell Kit (5-Prime; Hilden, Germany) including on-column DNase treatment, and quantified with the Epoch plate reader (Biotek; Winooski, Vermont, USA). For direct extraction from FACS-sorted cells, the RNeasy Micro Kit (Qiagen; Hilden, Germany) was used. The manufacturer's protocol was modified to allow direct extraction from collected cells. Briefly, cells were collected in 500 µL buffer RLT. Excess volume obtained from FACS was measured with a micropipette. Additional RLT was added to the sample to obtain a 350 µL to 100 µL ratio of RLT to excess liquid. For each 450 µL of total sample, 250 µL of 100% ethanol was added in place of the 350 µL of 70% ethanol typically used. The sample was loaded through the spin column in 700 µL stages, and the remainder of the protocol was performed unmodified, and included on-column DNase treatment.

3.4.2 Complementary DNA (cDNA) synthesis

Total RNA was reverse transcribed with the high capacity cDNA reverse transcription kit (Applied Biosystems; Foster City, California, USA) following manufacturer's instructions. Unless otherwise indicated, one (1) µg of total RNA was used in each reaction, and random hexamers were used to prime synthesis by reverse transcriptase. Reactions where oligo dT was used in place of random hexameters are indicated. Reactions performed with SuperScript IV Reverse Transcriptase (Thermo Fisher; Waltham, Massachusetts, USA) are indicated. "No RT" reactions included RNA template and all components except reverse transcriptase enzyme.

3.4.3 End-point PCR and sequencing

Primers for end-point PCR were designed using NCBI's Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast/). AmpliTaq Gold 360 Master Mix (Applied Biosystems) was used for all end-point PCR analyses. Primers were used at a final concentration of 250 nM. Cycling conditions were as follows:

1) Activation	95° C	5 min
2) Melting	95°C	30 sec
3) Annealing	Variable	30 sec
4) Extension	72°C	Variable. Return to step 2 for 35 total cycles.
5) Final Extension	72°C	10 min

Variable temperatures and times are indicated for each primer set, as are sequences of forward ("F") and reverse ("R") primers. Products were analyzed using agarose gel electrophoresis and band sizes were confirmed using the 1 kb plus or 100 bp plus DNA ladders (Thermo Fisher). All end-point PCR products were cloned into the pCR 4-TOPO plasmid with TOPO TA cloning for sequencing kit (Thermo Fisher). Cloning reactions were transformed into One Shot TOP10 Chemically Competent E. coli (Thermo Fisher). Individual clones were selected with Kanamycin and propagated for mini prep of plasmid DNA using the High-Speed Plasmid Mini Kit (Geneaid/FroggaBio; Toronto, Ontario, Canada). Multiple clones were sequenced for each product to confirm amplicon identities. Sanger sequencing using the plasmid-specific M13R or M13F primers was

conducted by the Molecular Biology Service Unit at the University of Alberta, or the London Regional Genomics Centre at Western University.

3.4.4 Exon junction end-point PCR

Primers used for Figure 3.1B and D:

set 1 F: GCCCTGCCCTGCTGTCCAAG

set 1 R: GGCTTGGCATGGAGGATATATTGCA

set 2 F: GTGGGGCAAGAGGCACCGTC

set 2 R: AGGCTTGGCATGGAGGATATATTGC

set 3 F: CTGCCCTGCTGTCCAAGGTCAT

set 3 R: ACTCGGTGGGGCTGGTAACG

ORF F: TATATAGCGATCGCCATGCACGCCCACTGCCTGCC

ORF R: ATATATACGCGTGCAGACTCTGAGGCTTGGCATGG

3.4.5 *NODAL* natural antisense transcript (NAT) PCR

The *NODAL* NAT was amplified from H9 and CA1 hES polyA+ cDNA with the following primers used for end-point PCR in Figure 3.11B:

F: GCAAGAGCTATGGTGGTTGTG

R: TAGCAAAGCTAGAGCCCTGTC

Annealing temperature: 54°C

Extension time: 2 minutes

3.4.6 Circular RNA PCR

Two pairs of non-overlapping primers were employed for *NODAL* exon 2 divergent PCR. Forward/ sense primers were designed near the 3' end of exon 2, and reverse/ antisense primers were designed near the 5' end of exon 2. Separate reactions were prepared for each set of primers.

Primers used for Figure 3.14B:

NODAL divergent exon2 F1 TACCCCAAGCAGTACAACGC

NODAL divergent exon2 R1 GTCCAGTTCTGCCCATCCAC

NODAL divergent exon2 F2		GTGAGGGCGAGTGTCCTAATC
NODAL divergent exon2 R2		TTGGCTCAGGAAGGAGAAGTC
Conditions:	Annealing temperature:	55°C.
	Extension time:	1 minute

3.4.7 3' Rapid Amplification of cDNA Ends (RACE) analyses

For 3' RACE, 2 µg total RNA was used for reverse transcription. Random primers were substituted for an oligo dT-adapter mix of "lock-dock" [63] primers with either A, G, or C as the most 3' base:

dT adapter primer R A: GGCCACGCGTCGACTAGTACTTTTTTTTTTTTTTTA

dT adapter primer R G: GGCCACGCGTCGACTAGTACTTTTTTTTTTTTTTTG

dT adapter primer R C: GGCCACGCGTCGACTAGTACTTTTTTTTTTTTTTTC

Each primer was used at a final concentration of 167 nM for a total primer concentration of 500 nM. 2 µL (equivalent to 200 ng RNA) of each cDNA reaction was used for subsequent PCR.

For the first round of amplification, primers were used at a final concentration of 200 nM:

Forward primers (variable for each analysis):

total *NODAL* 3' RACE F1:	TCTCCAAAGTAGTCTGTGTGTGAC
NODAL variant 3' RACE F1:	CTGCTGTCCAAGGTCATATGGG
NAT 3' RACE F1:	CGCTTCAGCCACTTGGAGAG

Reverse primer (identical for each analysis):

Abridged universal amplification primer (AUAP) R: GGCCACGCGTCGACTAGTAC

Conditions:

Annealing temperature:	54°C
Extension time:	2 minutes (total *NODAL*), 3 minutes (*NODAL* variant)

For the second (nested) round of amplification, 1 µL of PCR product from the first round of PCR was diluted into a final reaction volume of 20 µL for the nested PCR reaction conducted with the same conditions as round one, with the following primers:

Forward primers (variable for each analysis):

total *NODAL* 3'RACE F2 nested: TCCCCCTCCCCAAAGATTAAGG

NODAL variant 3' RACE F2 nested: AATATATCCTCCATGCCAAGCCTC

NAT 3' RACE F2 nested: ACCTCCAAAACCATGCTGCC

Reverse primer (identical for each analysis):

Abridged universal amplification primer (AUAP) R: GGCCACGCGTCGACTAGTAC

3.4.8 5' Rapid Amplification of cDNA Ends (RACE) analyses

5' RACE analysis was conducted using the 5' RACE System for Rapid Amplification of cDNA Ends (Thermo Fisher) following manufacturer's instructions. Three (3) µg of total RNA was used for each sample. Reverse transcription was performed for 50 minutes. cDNA (16.5 µL) was used for the tailing reaction. Tailed cDNA (2.5 µL) was used in a total volume of 25 µL for first round PCR. Primers were designed according to manufacturer's guidelines and to have melting temperatures similar to primers provided for a positive control target. "GSP" = gene-specific primer. All primers for first and second round PCR were used at a final concentration of 400 nM. 2.5 µL of 1/10 diluted first round PCR product was used for second round nested PCR.

Primers used for reverse transcription:

total *NODAL* 5' RACE GSP1 GAAAATCTCAATGGCAAGTGAG

NODAL variant 5' RACE GSP1 CATGGAGGATATATTGCAAGTC

Primers used for first round PCR:

total *NODAL* 5' RACE GSP2 CCATGCCAGATCCTCTTGTTG

NODAL variant 5' RACE GSP2 TCCCATATGACCTTGGACAGC

Abridged anchor primer (AAP) GGCCACGCGTCGACTAGTACGGGIIGGGIIGGGIIG

The same primer targeting constitutive exon 2 of *NODAL* was used for second round nested PCR analysis of both total *NODAL* and *NODAL* variant transcripts:

total *NODAL* 5' RACE nested GAAGGAGAAGTCAAAAGCAAACG

Abridged universal amplification primer (AUAP): GGCCACGCGTCGACTAGTAC

Conditions used:

Annealing temperature: 56°C

Extension time: 2 minutes

3.4.9 SYBR green real time PCR

Amplification of both *NODAL* isoforms in Figure 3.4A were amplified using the following primers using an annealing temperature of 60°C.

exon 2 F: TGTGAGGGCGAGTGTCC

exon 2 R: GAGGCACCCACATTCTTCCA

SYBR green real time PCR was performed using SsoAdvanced Universal SYBR Green Supermix (Bio-Rad; Hercules, California, USA). Primers were used at a final concentration of 100 nM. The following primers were used for Figure 3.4B-D.

SYBR green constitutive *NODAL* F: TACATCCAGAGTCTGCTG

SYBR green constitutive *NODAL* R: CCTTACTGGATTAGATGGTT

SYBR green *NODAL* variant F: CTGTTGGGGAGGAGTTTCA

SYBR green *NODAL* variant R: AGGCTTGGCATGGAGGATA

Cycling was performed on a CFX96 real time PCR detection system (Bio-Rad) using the following conditions:

1) Activation	95° C	10 min
2) Melting	95°C	15 sec
3) Annealing/ extension	60°C	1 min. Return to step 2 for 40 total cycles.

Results were analyzed using CFX manager (Bio-Rad) including a melt curve analysis to check for non-specific amplification. All SYBR green products were cloned and sequenced as described for end-point PCR amplicons. Cloned products for both *NODAL* and the *NODAL* variant were quantified using spectrophotometry and standard curves were prepared by calculating volumes required for a given number of target molecules.

3.4.10 Taqman real time PCR

Real time PCR was performed using Taqman gene expression master mix (Applied Biosystems) and Taqman gene expression assays for *POU5F1*/ *OCT4* (Hs04260367_gH), *NANOG* (Hs04260366_g1), *SOX2* (Hs01053049_s1), *RPLP0* (4333761), and *TBP*

(Hs99999910_m1). Expression was normalized to both *RPLP0* and *TBP* using the ΔΔCt method.

3.4.11 Duplexed *NODAL* splice variant ddPCR assay

Primers and probes for a digital droplet PCR assay to detect *NODAL* transcript isoforms were designed using primer 3 plus (http://primer3plus.com/). The following primers and probes were used, with fluorophores, internal quenchers, and terminal quenchers flanked by forward slashes.

Forward primer: GACCAACCATGCATACATC

Reverse primer: AACAAGTGGAAGGGACTC

Alternative exon probe:

/56-FAM/CCTGCTGTC/ZEN/CAAGGTCATAT/3IABkFQ/

Constitutive exon probe:

/5HEX/CTGGTAACG/ZEN/TTTCAGCAGAC/3IABkFQ/

Primers were used at a final concentration of 900 nM and probes were used at a final concentration of 250 nM. Droplets were generated and subject to a "two-step" PCR with the following conditions:

1) 95° C 10 min

2) 94°C 30 sec

3) 50°C 1 min

4) 72°C 2 min. Return to step 2 for 40 total cycles.

5) 98°C 10 min

Droplets that were both FAM-positive and HEX-positive, corresponding to the *NODAL* variant, were quantified using the Quantasoft software. Since constitutive *NODAL* was FAM-negative and HEX-positive, and could therefore be co-amplified in droplets containing *NODAL* variant transcript, constitutive *NODAL* was calculated manually using the equation: copies/ 20 µL sample = -ln(1-p) x 20,000 / 0.85. where 'p' is the proportion of positive droplets defined as FAM-HEX+ droplets / (FAM-HEX+ droplets + empty

droplets), and 0.85 nL is the average volume of a droplet as used by QuantaSoft (Bio-rad) [64].

3.4.12 Other ddPCR assays

Droplet digital PCR for total *NODAL* was conducted using Taqman primer probe assays Hs00415443_m1, Hs00250630_s1, or Hs01086749_m1 (Applied Biosystems). Unless indicated, Hs00415443_m1 was used for all ddPCR detection of total *NODAL* transcript. Primer probe assays were used at 1X (1/20th of supplied) concentration. For ddPCR detection of the *NODAL* NAT transcript, the following primers and probe were used at 900 nM and 250 nM, respectively.

NODAL NAT F	TTAATAGCAAAGCTAGAGCC
NODAL NAT R	CATGCATACATCCAGGTG
NODAL NAT FAM	/56-FAM/CCCAAGGCC/ZEN/AGCTTACTG/3IABkFQ/

The following cycling conditions were used:

1) 95° C	10 min
2) 94°C	30 sec
3) 55°C	1 min
5) 98°C	10 min

The number of target molecules detected was calculated using Quantasoft (Bio-Rad). For every sample, ddPCR was also used for detection of the housekeeping gene *RPLP0* using Taqman gene expression assay 4333761 (Applied Biosystems).

3.4.13 Microscopy

Pictures of H9 hES cells in different media were taken using EVOS FL Cell Imaging System (Thermo Fisher) with either 4X, 10X, or 20X objective lenses. Contrast and other image properties were adjusted so that cells and colony boundaries were more easily visible.

3.4.14 RNA stability experiments

H9 human embryonic stem cells grown in feeder-free conditions were treated for two or six hours with actinomycin D (Sigma-Aldrich; St. Louis, Missouri, USA). RNA extraction and was performed as described above. Equal volumes of RNA sample were reverse transcribed for each sample. Real time PCR was conducted in duplicate for the short half-life controls c-Myc (*MYC*) and TATA-binding protein (*TBP*; Hs99999910_m1), and for the long half-life control beta-Actin (*ACTB*; Hs01060665g1). A standard curve for each assay was used for target quantification in each sample. Detection of *NODAL* transcript isoforms was performed using the duplexed droplet digital PCR *NODAL* assay for *NODAL* splice variants. For each sample, 1 µL of cDNA was analyzed in duplicate or triplicate. Expression levels for each transcript of interest for each sample were normalized to *ACTB* levels and to the average transcript level within each experiment. Expression for each experiment was reported relative to cells that did not receive any actinomycin D treatment (t = 0). Each target of interest was fitted by an exponential trend line in Microsoft Excel for Mac (version 15.4, Microsoft), so that half-lives could be calculated based on the returned equation in the form: $N(t) = N_0 e^{-\lambda t}$, where $N(t)$ is the quantity at a given time t. N_0 is the quantity at t=0, and λ is the exponential decay constant. Thus the half-life can be calculated using $t_{1/2} = \ln(2)/-\lambda$. For comparisons of the half-lives of two different transcripts, a one-way analysis of covariance (ANCOVA) for independent samples was conducted on log-transformed relative expression values for all actinomycin D-treated samples using treatment time as the concomitant variable and performed using Vassar Stats (http://vassarstats.net/vsancova.html).

3.4.15 Morpholino experiments

Antisense morpholino oligonucleotides (MOs) were synthesized by Gene Tools. All MOs had a 3' fluorescein tag. The "standard control oligo" was used in all control treatments. MOs with the following sequences were used to target *NODAL*:

NODAL alternative exon (SNP T)	AGACCCTGAATCCCACCTGAGGCTT
NODAL constitutive exon 2	CCTCACGCCTGGCATCCCACCTGGA

H9 hES cells were grown in feeder-free conditions on Matrigel and with MEF-CM. When ready for passage, cells were treated with 20 μM MO. In the presence of MO, colonies were manually passaged and transferred to a new culture vessel at a 1:2 split ratio. MO-containing media was replaced 18 hours later with fresh media that did not contain MO. After 48 hours, cells were sorted using FACS within the Faculty of Medicine and Dentistry Flow Cytometry Facility at the University of Alberta. The top 25-50% of fluorescein-positive cells for each treatment were collected for direct isolation of total RNA as described above. In subsequent PCR assays, expression of each sample was normalized using Taqman gene expression assays (Applied Biosystems) for three housekeeping genes *RPLPO* (4333761), *TBP* (Hs99999910_m1), and 18S (4333760F).

3.4.16 PCR arrays

The human "Embryonic Stem Cells" RT^2 Profiler PCR Array (SA Biosciences/ Qiagen) was used for SYBR green real time PCR detection of genes related to human embryonic stem cell pluripotency and differentiation. Plates were cycled according to manufacturer's instructions using the CFX 96 real time PCR system and results were analyzed with CFX manager (Bio-Rad). Melt curve analysis was used to exclude samples with low melt peaks and inconsistent melt profiles for the same target between samples, indicative of non-specific amplification. Genes with any excluded samples were not included in heat maps comparing the effects of *NODAL* and *NODAL* variant knockdowns. Expression values were normalized using the median for all five endogenous control targets included in the array, and the ΔΔCt method.

3.5 References

1. de Klerk, E., & t Hoen, P. A. C. (2015). Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends in Genetics*, *31*(3), 128–139. doi:10.1016/j.tig.2015.01.001
2. Khorkova, O., Myers, A. J., Hsiao, J., & Wahlestedt, C. (2014). Natural antisense transcripts. *Human Molecular Genetics*, *23*(R1), R54–R63. doi:10.1093/hmg/ddu207
3. Wight, M., & Werner, A. (2013). The functions of natural antisense transcripts. *Essays In Biochemistry*, *54*, 91–101. doi:10.1042/bse0540091

4. Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. doi:10.1038/nature07509

5. Yeo, G. W., Xu, X., Liang, T. Y., Muotri, A. R., Carson, C. T., Coufal, N. G., & Gage, F. H. (2007). Alternative Splicing Events Identified in Human Embryonic Stem Cells and Neural Progenitors. *PLoS computational biology*, *3*(10), e196–17. doi:10.1371/journal.pcbi.0030196

6. Salomonis, N., Nelson, B., Vranizan, K., Pico, A. R., Hanspers, K., Kuchinsky, A., et al. (2009). Alternative Splicing in the Differentiation of Human Embryonic Stem Cells into Cardiac Precursors. *PLoS computational biology*, *5*(11), e1000553–17. doi:10.1371/journal.pcbi.1000553

7. David, C. J., & Manley, J. L. (2010). Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes & Development*, *24*(21), 2343–2364. doi:10.1101/gad.1973010

8. Fackenthal, J. D., & Godley, L. A. (2008). Aberrant RNA splicing and its functional consequences in cancer cells. *Disease Models and Mechanisms*, *1*(1), 37–42. doi:10.1242/dmm.000331

9. Srebrow, A., & Kornblihtt, A. R. (2006). The connection between splicing and cancer. *Journal of Cell Science*, *119*(Pt 13), 2635–2641. doi:10.1242/jcs.03053

10. Holm, F., Hellqvist, E., Mason, C. N., Ali, S. A., Delos-Santos, N., Barrett, C. L., et al. (2015). Reversion to an embryonic alternative splicing program enhances leukemia stem cell self-renewal. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(50), 15444–15449. doi:10.1073/pnas.1506943112

11. Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., et al. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA*, *19*(2), 141–157. doi:10.1261/rna.035667.112

12. Tian, B. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*, *33*(1), 201–212. doi:10.1093/nar/gki158

13. Derti, A., Garrett-Engele, P., MacIsaac, K. D., Stevens, R. C., Sriram, S., Chen, R., et al. (2012). A quantitative atlas of polyadenylation in five mammals. *Genome research*, *22*(6), 1173–1183. doi:10.1101/gr.132563.111

14. Tian, B., & Manley, J. L. (2016). Alternative polyadenylation of mRNA precursors. *Nature reviews. Molecular cell biology*, 1–13. doi:10.1038/nrm.2016.116

15. Lackford, B., Yao, C., Charles, G. M., Weng, L., Zheng, X., Choi, E. A., et al.

(2014). Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. *The EMBO Journal*, *33*(8), 878–889. doi:10.1002/embj.201386537

16. Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., et al. (2012). Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nature Methods*, *10*(2), 133–139. doi:10.1038/nmeth.2288

17. Elkon, R., Ugalde, A. P., & Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. *Nature Reviews Genetics*, *14*(7), 496–506. doi:10.1038/nrg3482

18. Movassat, M., Crabb, T. L., Busch, A., Yao, C., Reynolds, D. J., Shi, Y., & Hertel, K. J. (2016). Coupling between alternative polyadenylation and alternative splicing is limited to terminal introns. *RNA biology*, *13*(7), 646–655. doi:10.1080/15476286.2016.1191727

19. Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., et al. (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*, *164*(4), 805–817. doi:10.1016/j.cell.2016.01.029

20. Saijoh, Y., Oki, S., Tanaka, C., Nakamura, T., Adachi, H., Yan, Y.-T., et al. (2005). Two nodal-responsive enhancers control left-right asymmetric expression ofNodal. *Developmental Dynamics*, *232*(4), 1031–1036. doi:10.1002/dvdy.20192

21. Papanayotou, C., Benhaddou, A., Camus, A., Perea-Gomez, A., Jouneau, A., Mezger, V., et al. (2014). A Novel Nodal Enhancer Dependent on Pluripotency Factors and Smad2/3 Signaling Conditions a Regulatory Switch During Epiblast Maturation. *PLoS Biology*, *12*(6), e1001890–14. doi:10.1371/journal.pbio.1001890

22. Vincent, S. D., Norris, D. P., Ann Le Good, J., Constam, D. B., & Robertson, E. J. (2004). Asymmetric Nodal expression in the mouse is governed by the combinatorial activities of two distinct regulatory elements. *Mechanisms of Development*, *121*(11), 1403–1415. doi:10.1016/j.mod.2004.06.002

23. Norris, D. P., & Robertson, E. J. (1999). Asymmetric and node-specific nodal expression patterns are controlled by two distinct cis-acting regulatory elements. *Genes & Development*, *13*(12), 1575–1588.

24. Strizzi, L., Hardy, K. M., Kirschmann, D. A., Ahrlund-Richter, L., & Hendrix, M. J. C. (2012). Nodal Expression and Detection in Cancer: Experience and Challenges. *Cancer Research*, *72*(8), 1915–1920. doi:10.1158/0008-5472.CAN-11-3419

25. Proudfoot, N. J. (2011). Ending the message: poly(A) signals then and now. *Genes & Development*, *25*(17), 1770–1782. doi:10.1101/gad.17268411

26. Retelska, D., Iseli, C., Bucher, P., Jongeneel, C. V., & Naef, F. (2006). BMC Genomics. *BMC Genomics*, *7*(1), 176–10. doi:10.1186/1471-2164-7-176

27. Dutertre, M., Sanchez, G., De Cian, M.-C., Barbier, J., Dardenne, E., Gratadou, L., et al. (2010). Cotranscriptional exon skipping in the genotoxic stress response. *Nature Structural & Molecular Biology*, *17*(11), 1358–1366. doi:10.1038/nsmb.1912

28. Walton, H. S., Gebhardt, F. M., Innes, D. J., & Dodd, P. R. (2007). Analysis of multiple exon-skipping mRNA splice variants using SYBR Green real-time RT-PCR. *Journal of Neuroscience Methods*, *160*(2), 294–301. doi:10.1016/j.jneumeth.2006.09.022

29. Quail, D. F., Siegers, G. M., Jewer, M., & Postovit, L.-M. (2013). Nodal signalling in embryogenesis and tumourigenesis. *The International Journal of Biochemistry & Cell Biology*, *45*(4), 885–898. doi:10.1016/j.biocel.2012.12.021

30. Bodenstine, T. M., Chandler, G. S., Seftor, R. E. B., Seftor, E. A., & Hendrix, M. J. C. (2016). Plasticity underlies tumor progression: role of Nodal signaling. *Cancer and Metastasis Reviews*, *35*(1), 21–39. doi:10.1007/s10555-016-9605-5

31. Quail, D. F., Zhang, G., Walsh, L. A., Siegers, G. M., Dieters-Castator, D. Z., Findlay, S. D., et al. (2012). Embryonic Morphogen Nodal Promotes Breast Cancer Growth and Progression. *PLoS ONE*, *7*(11), e48237–12. doi:10.1371/journal.pone.0048237

32. Kirsammer, G., Strizzi, L., Margaryan, N. V., Gilgur, A., Hyser, M., Atkinson, J., et al. (2014). Nodal signaling promotes a tumorigenic phenotype in human breast cancer. *Seminars in Cancer Biology*, *29*, 40–50. doi:10.1016/j.semcancer.2014.07.007

33. Topczewska, J. M., Postovit, L.-M., Margaryan, N. V., Sam, A., Hess, A. R., Wheaton, W. W., et al. (2006). Embryonic and tumorigenic pathways converge via Nodal signaling: role in melanoma aggressiveness. *Nature Medicine*, *12*(8), 925–932. doi:10.1038/nm1448

34. Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, *41*(D1), D36–D42. doi:10.1093/nar/gks1195

35. Sharova, L. V., Sharov, A. A., Nedorezov, T., Piao, Y., Shaik, N., & Ko, M. S. H. (2009). Database for mRNA Half-Life of 19 977 Genes Obtained by DNA Microarray Analysis of Pluripotent and Differentiating Mouse Embryonic Stem Cells. *DNA Research*, *16*(1), 45–58. doi:10.1093/dnares/dsn030

36. Huang, Y., Liang, P., Liu, D., Huang, J., & Songyang, Z. (2014). Telomere regulation in pluripotent stem cells. *Protein & Cell*, *5*(3), 194–202.

doi:10.1007/s13238-014-0028-1

37. Chappell, J., & Dalton, S. (2013). Roles for MYC in the establishment and maintenance of pluripotency. *Cold Spring Harbor Perspectives in Medicine*, *3*(12), a014381–a014381. doi:10.1101/cshperspect.a014381

38. Lindeboom, R. G. H., Supek, F., & Lehner, B. (2016). The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nature genetics*, *48*(10), 1112–1118. doi:10.1038/ng.3664

39. Lykke-Andersen, S., & Jensen, T. H. (2015). Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nature reviews. Molecular cell biology*, *16*(11), 665–677. doi:10.1038/nrm4063

40. Brogna, S., & Wen, J. (2009). Nonsense-mediated mRNA decay (NMD) mechanisms. *Nature Structural & Molecular Biology*, *16*(2), 107–113. doi:10.1038/nsmb.1550

41. Bates, D. O., Mavrou, A., Qiu, Y., Carter, J. G., Hamdollah-Zadeh, M., Barratt, S., et al. (2013). Detection of VEGF-Axxxb Isoforms in Human Tissues. *PLoS ONE*, *8*(7), e68399–10. doi:10.1371/journal.pone.0068399

42. Vallier, L., Mendjan, S., Brown, S., Chng, Z., Teo, A., Smithers, L. E., et al. (2009). Activin/Nodal signalling maintains pluripotency by controlling Nanog expression. *Development*, *136*(8), 1339–1349. doi:10.1242/dev.033951

43. Vallier, L. (2005). Activin/Nodal and FGF pathways cooperate to maintain pluripotency of human embryonic stem cells. *Journal of Cell Science*, *118*(19), 4495–4509. doi:10.1242/jcs.02553

44. Adewumi, O., Aflatoonian, B., Ahrlund-Richter, L., Amit, M., Andrews, P. W., Beighton, G., et al. (2007). Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. *Nature Biotechnology*, *25*(7), 803–816. doi:10.1038/nbt1318

45. Besser, D. (2004). Expression of Nodal, Lefty-A, and Lefty-B in Undifferentiated Human Embryonic Stem Cells Requires Activation of Smad2/3. *Journal of Biological Chemistry*, *279*(43), 45076–45084. doi:10.1074/jbc.M404979200

46. Inman, G. J., Nicolás, F. J., Callahan, J. F., Harling, J. D., Gaster, L. M., Reith, A. D., et al. (2002). SB-431542 is a potent and specific inhibitor of transforming growth factor-beta superfamily type I activin receptor-like kinase (ALK) receptors ALK4, ALK5, and ALK7. *Molecular pharmacology*, *62*(1), 65–74.

47. James, D. (2005). TGF /activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. *Development*, *132*(6), 1273–1282. doi:10.1242/dev.01706

48. Piccolo, S., Agius, E., Leyns, L., Bhattacharyya, S., Grunz, H., Bouwmeester, T., & De Robertis, E. M. (1999). The head inducer Cerberus is a multifunctional antagonist of Nodal, BMP and Wnt signals. *Nature*, *397*(6721), 707–710. doi:10.1038/17820

49. Wu, J., & Izpisua Belmonte, J. C. (2015). Dynamic Pluripotent Stem Cell States and Their Applications. *Cell Stem Cell*, *17*(5), 509–525. doi:10.1016/j.stem.2015.10.009

50. Ávila-González, D., García-López, G., García-Castro, I. L., Flores-Herrera, H., Molina-Hernández, A., Portillo, W., & Díaz, N. F. (2016). Capturing the ephemeral human pluripotent state. *Developmental Dynamics*, *245*(7), 762–773. doi:10.1002/dvdy.24405

51. Hayashi, Y., & Furue, M. K. (2016). Biological Effects of Culture Substrates on Human Pluripotent Stem Cells. *Stem Cells International*, *2016*, 1–11. doi:10.1155/2016/5380560

52. Hannoun, Z., Fletcher, J., Greenhough, S., Medine, C., Samuel, K., Sharma, R., et al. (2010). The Comparison between Conditioned Media and Serum-Free Media in Human Embryonic Stem Cell Culture and Differentiation. *Cellular Reprogramming (Formerly "Cloning and Stem Cells")*, *12*(2), 133–140. doi:10.1089/cell.2009.0099

53. Strizzi, L., Sandomenico, A., Margaryan, N. V., Focà, A., Sanguigno, L., Bodenstine, T. M., et al. (2015). Effects of a novel Nodal-targeting monoclonal antibody in melanoma. *Oncotarget*, *6*(33), 34071–34086. doi:10.18632/oncotarget.6049

54. Arai, D., Hayakawa, K., Ohgane, J., Hirosawa, M., Nakao, Y., Tanaka, S., & Shiota, K. (2015). An epigenetic regulatory element of the Nodal gene in the mouse and human genomes. *Mechanisms of Development*, *136*, 143–154. doi:10.1016/j.mod.2014.12.003

55. Sampath, K., & Robertson, E. J. (2016). Keeping a lid on nodal: transcriptional and translational repression of nodal signalling. *Open Biology*, *6*(1), 150200–8. doi:10.1098/rsob.150200

56. Costa, F. F., Seftor, E. A., Bischof, J. M., Kirschmann, D. A., Strizzi, L., Arndt, K., et al. (2009). Epigenetically reprogramming metastatic tumor cells with an embryonic microenvironment. *Epigenomics*, *1*(2), 387–398. doi:10.2217/epi.09.25

57. Hardy, K. M., Kirschmann, D. A., Seftor, E. A., Margaryan, N. V., Postovit, L. M., Strizzi, L., & Hendrix, M. J. C. (2010). Regulation of the Embryonic Morphogen Nodal by Notch4 Facilitates Manifestation of the Aggressive Melanoma Phenotype. *Cancer Research*, *70*(24), 10340–10350. doi:10.1158/0008-

5472.CAN-10-0705

58. Lawrence, M. G., Margaryan, N. V., Loessner, D., Collins, A., Kerr, K. M., Turner, M., et al. (2011). Reactivation of embryonic nodal signaling is associated with tumor progression and promotes the growth of prostate cancer cells. *The Prostate*, *71*(11), 1198–1209. doi:10.1002/pros.21335

59. Hardy, K. M., Strizzi, L., Margaryan, N. V., Gupta, K., Murphy, G. F., Scolyer, R. A., & Hendrix, M. J. C. (2015). Targeting Nodal in Conjunction with Dacarbazine Induces Synergistic Anticancer Effects in Metastatic Melanoma. *Molecular Cancer Research*, *13*(4), 670–680. doi:10.1158/1541-7786.MCR-14-0077

60. Postovit, L.-M., Margaryan, N. V., Seftor, E. A., Kirschmann, D. A., Lipavsky, A., Wheaton, W. W., et al. (2008). Human embryonic stem cell microenvironment suppresses the tumorigenic phenotype of aggressive cancer cells. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(11), 4329–4334. doi:10.1073/pnas.0800467105

61. Quail, D. F., Taylor, M. J., Walsh, L. A., Dieters-Castator, D., Das, P., Jewer, M., et al. (2011). Low oxygen levels induce the expression of the embryonic morphogen Nodal. *Molecular biology of the cell*, *22*(24), 4809–4821. doi:10.1091/mbc.E11-03-0263

62. Lonardo, E., Hermann, P. C., Mueller, M.-T., Huber, S., Balic, A., Miranda-Lorenzo, I., et al. (2011). Nodal/Activin Signaling Drives Self-Renewal and Tumorigenicity of Pancreatic Cancer Stem Cells and Provides a Target for Combined Drug Therapy. *Cell Stem Cell*, *9*(5), 433–446. doi:10.1016/j.stem.2011.10.001

63. Borson, N. D., Salo, W. L., & Drewes, L. R. (1992). A lock-docking oligo(dT) primer for 5“ and 3” RACE PCR. *PCR methods and applications*, *2*(2), 144–148.

64. Corbisier, P., Pinheiro, L., Mazoua, S., Kortekaas, A.-M., Chung, P. Y. J., Gerganova, T., et al. (2015). DNA copy number concentration measured by digital and droplet digital quantitative PCR using certified reference materials. *Analytical and Bioanalytical Chemistry*, *407*(7), 1831–1840. doi:10.1007/s00216-015-8458-z

Chapter 4

4 Function and post-translational regulation of NODAL proteins

4.1 Introduction

Nodal is a TGF-beta superfamily member with several key roles in early embryonic development in vertebrates. These include specification of mesendoderm, induction of gastrulation, and establishment of anterior-posterior and left-right axes (reviewed in [1-4]). There are several aspects to Nodal protein dynamics that are key to its function in the early embryo. As a paracrine growth factor, Nodal must be efficiently secreted from expressing cells to exert its effects on both neighbouring and distant cells as an embryonic morphogen. A 26 amino acid N-terminal signal peptide directs nascent translated peptide into the endoplasmic reticulum (ER) for processing along the secretory pathway. The remainder of the protein is translated into the ER as an approximately 37 kDa pro-protein consisting of a 211-amino acid N-terminal pro-domain of about 24 kDa, and a 110-amino acid C-terminal mature domain of about 13 kDa.

Cleavage of the pro-domain from the mature Nodal peptide occurs as a result of the proteolytic activity of secreted pro-protein convertases Furin and Pcsk6 (also known as Pace4) via recognition of an R-X-R-R motif [5, 6]. The resultant mature Nodal peptides can homo-dimerize to engage both type I tyrosine kinase receptors Alk4 or Alk7 (also known as Acvr1B and Acvr1C, respectively), and type II receptors Acvr2A or Acvr2B (formerly known as ActrIIa and ActRIIB, respectively). Receptor complex formation triggers phosphorylation of mediator Smads (Smad2 and Smad3) and their subsequent interaction with Smad4 to affect expression of target genes (reviewed in [7]).

Proteolytic processing enhances or activates Nodal signalling, as it is a point of regulation for Nodal signalling range in the embryo [8]. Furthermore, processing is essential for the induction of mesendoderm differentiation and subsequent gastrulation of the mouse epiblast [6, 9]. However, mutant *Nodal* (*Xnr2*) constructs resistant to cleavage were still able to induce mesoderm in xenopus [10], which is induced at a lower threshold of signal

relative to endoderm [1]. Also, homozygous cleavage-resistant *Nodal* mice still underwent EMT and primitive streak formation before development was arrested, whereas *Nodal*-null mice lacked a primitive streak entirely [11]. Several dominant negative roles for cleavage-resistant Nodals have also been described [12]. Collectively, these results suggest that proteolytic processing plays an important role in the regulation of Nodal activity.

Two GPI-linked and membrane bound members of the epidermal growth factor-cysteine-rich Cripto-1/FLR1/cryptic (EGF-CFC) family serve as co-receptors for Nodal signals. Cripto (Tdgf1) and Cryptic (Cfc1) bind the type I Nodal receptor and help recruit type II receptors to facilitate a functional receptor complex [13, 14]. Beyond receptor assembly, Cripto also binds Furin and Pace4 on signal-receiving cells to facilitate proteolytic maturation of Nodal [15]. Furthermore, Cripto facilitates Nodal inhibition by Lefty proteins, as direct binding of Lefty to either Nodal or Cripto/Cryptic can prevent successful receptor-ligand complex formation [16]. Interestingly, although Cripto is generally required for Nodal signalling, Cripto-independent signalling has been described in the mouse embryo [17, 18]. Collectively, these findings suggest a role for EGF-CFC proteins as multifaceted facilitators of Nodal signalling. Among TGF-beta superfamily members, Nodal is not unique in its utilization of EGF-CFC co-receptors, as growth and differentiation factors Gdf1 and Gdf3 also utilize the same receptors as Nodal [2]. Interestingly, Gdf1 can also hetero-dimerize with Nodal [19, 20].

Another TGF-beta superfamily member closely related to Nodal is Activin. While utilizing the same receptor complexes as Nodal and also signalling through Smad2 and Smad3, Activin signalling is distinct from Nodal in that it does not require Cripto as a co-receptor, and is refractory to inhibition by Lefty [21, 22]. However, Cripto can participate in Activin receptor complex formation, where it actually inhibits productive signal transduction [23].

Elegant work by Cheng and colleagues utilized chimeras of squint (*sqt*; a zebrafish *Nodal*) and Activin to determine specific regions of the mature peptide responsible for Nodal's Cripto dependence [22]. A construct where the most C-terminal third of the

Nodal mature domain corresponding to the second “finger” projection was replaced with Activin sequence was able to induce ectopic expression of both *gsc* and *ntl*, even in one-eyed pinhead (*oep*; a zebrafish Cripto homolog) mutant embryos, thus relinquishing Nodal’s dependence on Cripto. This work highlights the importance of the C-terminus of Nodal in conferring its function and specificity relative to other related ligands.

To investigate which structural aspects confer such divergent function between the closely related NODAL and Activin proteins, chimeras of NODAL and TGF-beta superfamily member BMP2 mature peptides were generated and screened for proteins that could both refold efficiently and induce NODAL phenotypes *in vitro* and *in vivo* [24]. One such chimera (NB250) consisting of N-terminal and C-terminal BMP2 sequence flanking a large segment of NODAL sequence, induced SMAD2 phosphorylation in cells over-expressing Cripto, and was able to reverse heart looping in chick embryos. A corresponding crystal structure for this chimera revealed a BMP2-like structure. The authors suggested this is evidence that NODAL likely folds in a similar fashion to BMP2, although it should be noted that despite the NODAL functionality of the chimera, the C-terminal region known to confer functional specificity to NODAL did consist of BMP2 sequence.

A major structural and functional characteristic of TGF-beta superfamily members is their ability to form intrachain and interchain disulfide bonds. Indeed, NODAL contains a set of seven cysteines in its mature domain analogous to other TGF-beta superfamily members, with six of these cysteines participating in intrachain disulfide bonds, and the seventh cysteine participating in an interchain disulfide bond to form a Nodal-Nodal homodimer. While highly conserved across superfamily members both within and between vertebrate species, these cysteines, and their corresponding disulfide bonds and homo-dimerization have not been directly experimentally assessed for human NODAL.

One post-translational modification characteristic of TGF-beta superfamily members and secreted proteins in general is N-glycosylation, which consists of the covalent addition of a glycan oligosaccharide to asparagine residues within N-X-S/T motifs [25]. Full-length intracellular pro-Nodal is found in an N-glycosylated form [26], and corresponding pro-

Nodal secreted into conditioned media was found to contain complex carbohydrate modifications, indicative of further N-glycan processing along the secretory pathway. Similar modifications to both full-length pro-Nodal and the cleaved pro-domain indicate that the pro-domain is the site of these post-translational modifications, although the exact sites of these modifications were not assessed. N-glycosylation generally aids in protein folding in the ER and thus protein stability (reviewed in [27]). In contrast to the pro-domain, the mature peptides of both human and mouse mature NODAL/Nodal ligands do not contain N-glycosylation motifs. Once cleaved from the N-glycosylated pro-domain, it has been suggested that the mature Nodal peptide is rapidly degraded and thus limited in its signalling range [28]. Interestingly, experimental introduction of different N-glycosylation motifs found in Bmp6 or the Xenopus nodal related (Xnr) proteins into the Nodal mature domain increased the accumulation of mature Nodal peptide in conditioned media and consequently signalling range in zebrafish blastulae [28]. However, the effect of this N-glycosylation on Nodal secretion, processing, or dimerization was not reported.

Multiple functional *NODAL*-related genes are present in the genomes of some model organisms such as zebrafish (squint, cyclops, and southpaw), and xenopus (Xnr1-6). Conversely, mouse and human each have only one *NODAL/Nodal* gene. To date, no proteins resulting from alternative splicing or otherwise differentially processed transcripts have been described for human *NODAL* or mouse *Nodal*. Furthermore, much of our knowledge of NODAL protein function is provided by study of endogenous NODAL-related genes in non-human systems. While these studies have made immense and important contributions to our understanding of NODAL biology and many NODAL functions are highly conserved between species, there is a distinct lack of work characterizing the processing and dynamics of human NODAL protein specifically. Appreciation of nuances between species will undoubtedly help improve modelling of human NODAL function. This is paramount for the advancement of regenerative medicine projects and cancer therapy development where human specific NODAL biology is highly relevant.

In recent years it has become increasingly appreciated that a large number of protein coding genes are subject to at least some degree of alternative splicing, a process found to be much more widespread in primates including humans relative to other vertebrates [29]. It is now estimated that over 90% of multi-exon protein coding gene loci are subject to alternative splicing [30, 31]. Not all alternatively spliced RNA transcripts are ultimately translated into protein. For example, some alternative splicing events have been shown to introduce premature termination codons (PTCs) as a means of negatively regulating gene expression through nonsense-mediated decay (NMD) of affected transcripts [32-35]. Still, alternative splicing is widely accepted as a major contributor to the generation of proteomic diversity from a limited genome [36]. It has been recently proposed that different distinct proteins coded by the same gene locus be identified as "proteoforms" analogous to the term "isoform" used to describe distinct nucleic acids from the same gene [37]. To date, the extent to which alternative splicing contributes to productive translation of multiple proteoforms from a single locus on a genome-wide scale remains unclear. For example, a comprehensive search of the protein data bank (PDB) revealed only 15 genes for which experimentally confirmed protein structures corresponding to translated products of multiple mRNA isoforms have been obtained [38], underscoring the dramatic lack of genome-wide characterization of alternative splicing at the protein level. Despite this, an increasing number of individually studied alternative splicing events illustrate cases of alternative splicing affecting protein localization, protein-protein interactions, protein domain structure, and protein stability, as well as enzymatic properties and other protein functions (reviewed in [39, 40]). Hence, alternative splicing likely contributes substantially to the production of a complex and functionally diverse proteome.

The discovery of a full-length processed splice variant for *NODAL* detailed in the previous chapters prompted my investigation of this novel isoform at the protein level. This chapter consists of a comprehensive comparative assessment of the two NODAL proteoforms in terms of their post-translational modification, secretion, proteolytic processing, extracellular dynamics, complex formation, and signalling capacity. Novel aspects of constitutive NODAL processing are also revealed that complement previous studies on the topic.

4.2 Results

Alternative splicing of human *NODAL* results in inclusion of a 116 base-pair cassette exon downstream of constitutive exon 2 that codes for unique amino acids. Inclusion of this alternative exon also alters the translational reading frame, resulting in non-constitutive NODAL peptide sequence into constitutive exon 3. Shortly into constitutive exon 3, this altered translational reading frame results in a TGA "stop" codon marking the end of a 338 amino acid open reading frame (ORF), relative to constitutive NODAL's 347. The NODAL variant and constitutive NODAL proteins share identical signal peptides, pro-domains, and N-terminal halves of the NODAL mature peptide. The constitutive C-terminal NODAL sequence is absent in the NODAL variant protein, where 41 unique amino acids are instead found (Figure 4.1). A sequence alignment between the mature domains of constitutive NODAL and the NODAL variant reveals partial alignment in the divergent C-terminal region (Figure 4.2). Overall, these domains share identical amino acids at 55% of the alignment positions. Downstream of the amino acids coded by constitutive exon 2, the NODAL proteoforms are distinct in sequence, with alignment indicating identical amino acids at 14% of positions and similar amino acids at 17% of positions (Figure 4.2). The unique C-terminal NODAL variant sequence did not contain any known protein domains and did not return any BLAST alignments with E-values of less than 1.

I used two general approaches to compare the two NODAL proteoforms. First, sequence-based approaches were used to assess potential differences in domain structure and sites of post-translational modification. Second, analysis of previously reported experimentally generated structures as well as structural prediction models were used to compare potential structural differences between the two NODAL proteoforms. Results of these analyses were incorporated into experimental modelling.

For direct experimental study of NODAL proteins, I generated expression vectors for each NODAL proteoform with C-terminal MYC-DYK tags. These constructs were used for over-expression in HEK 293 cells. Western blot analysis of cell lysates revealed multiple bands for each NODAL proteoform. Specifically, constitutive NODAL

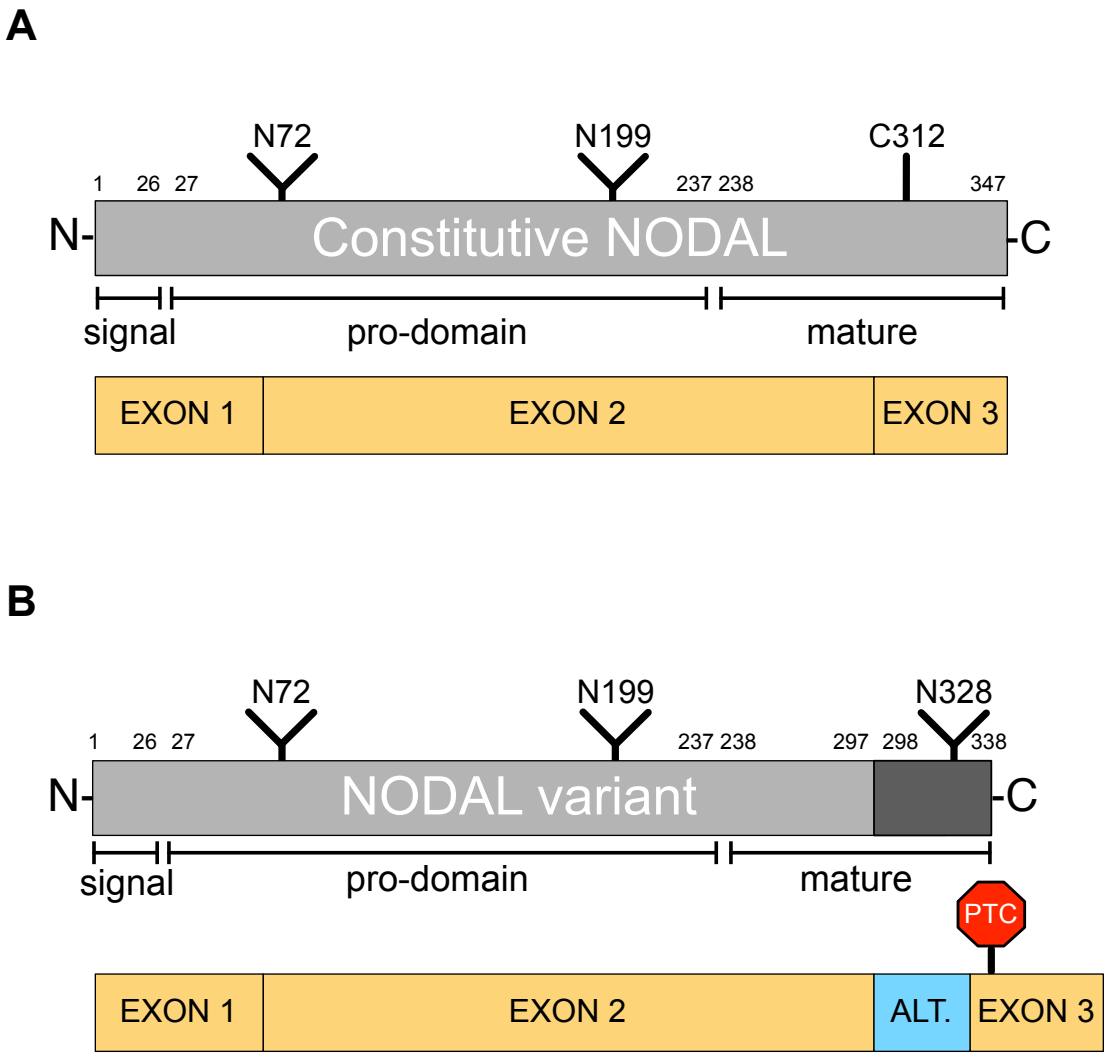


Figure 4.1: Constitutive *NODAL* and the *NODAL* variant open reading frames differ in sequence at the C-terminal region of the mature NODAL peptide.

Proteins are shown with N-terminus at the left and C-terminus at the right. Numbers mark amino acid positions for the start and end of each element. N72, N199, and N328 mark positions of putative N-glycosylation sites. A) Constitutive *NODAL* open reading frame. C312 marks cysteine at position 312 involved in putative interchain disulfide bond formation. B) The *NODAL* variant open reading frame. The darker protein region indicates novel sequence unique from constitutive NODAL. “PTC” = premature termination codon. “ALT.” = alternatively spliced cassette exon.

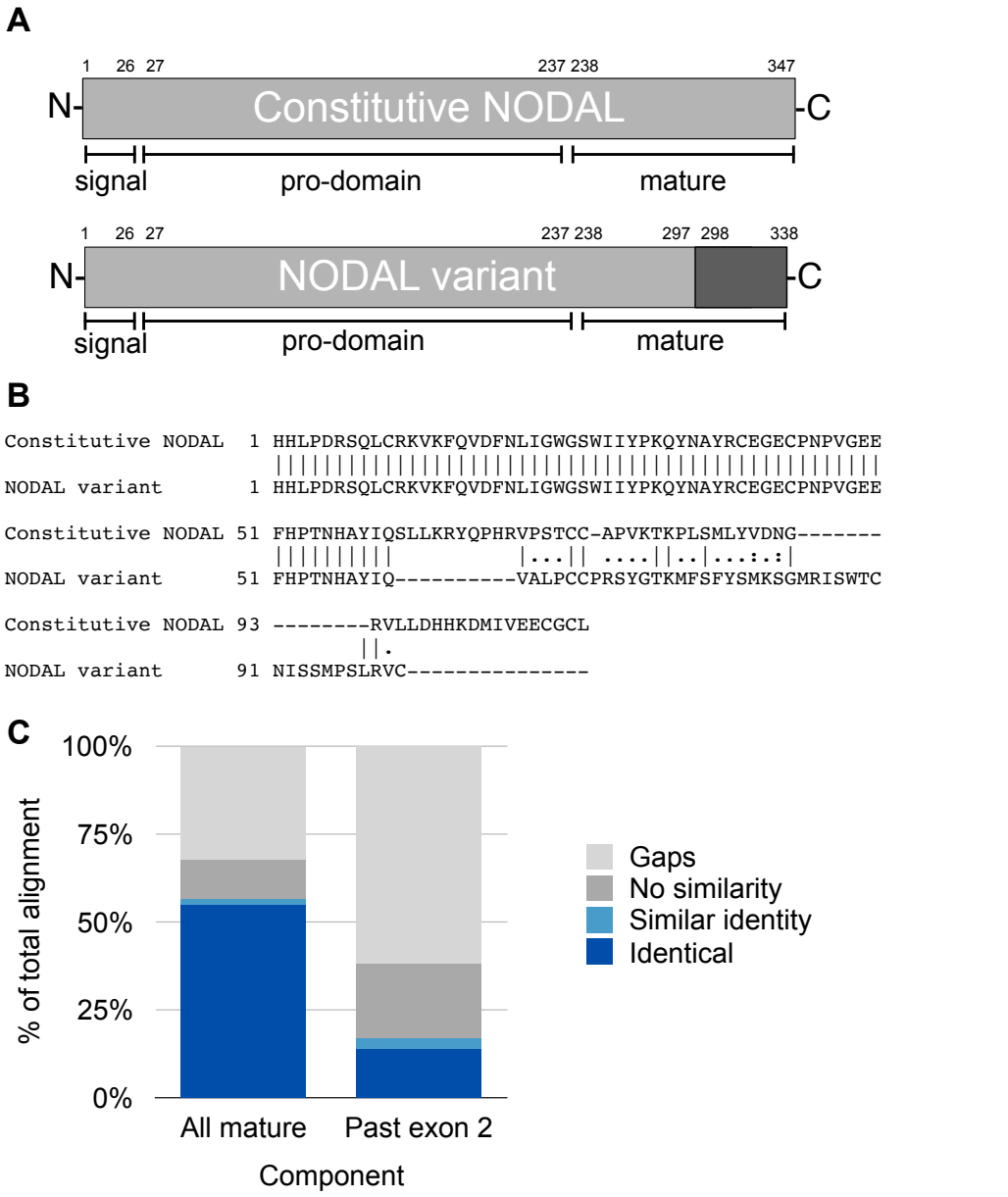


Figure 4.2: Sequence alignment between constitutive NODAL and the NODAL variant proteins.

A) Darker region of the NODAL variant indicates unique peptide sequence from constitutive NODAL. B) EMBOSS Needle pairwise alignment between constitutive NODAL and the NODAL variant mature peptides only. Numbers indicate position in the mature peptide from N-terminus to C-terminus. "|" = exact match amino acid pairs. ":" = similar amino acids. "." = no similarity. "-" = gap in alignment. C) Aligned amino acids by category. "All mature" = entire alignment from B. "Past exon 2" = aligned amino acids coded by the *NODAL* variant open reading frame downstream of exon 2 which is common to both isoforms.

expression resulted in two bands very close in size and consistent with the predicted size of full-length protein. NODAL variant expression resulted in three bands very close in size and similar in size to those for NODAL (Figure 4.3).

Since NODAL is putatively N-glycosylated and N-glycosylation sites have a well-defined (although not deterministic) N-X-S/T-X (where X is any amino acid other than proline) sequence motif, I was interested in determining whether the different banding pattern between the two proteoforms was the result of differential N-glycosylation. Furthermore, N-glycosylation is known to be important for secreted protein function, and has been artificially shown to enhance NODAL signalling range in the zebrafish embryo [28]. However, the endogenous N-glycosylation of NODAL has not been directly explored. Indeed, N-glycosylation site prediction using the NetNGlyc tool revealed two N-glycosylation motifs in the pro-domain shared by both NODAL proteoforms, and a third unique potential N-glycosylation site in the mature domain of the NODAL variant. The most N-terminal N-glycosylation motif at N72 was predicted to be unmodified, while the motif at N199 and the motif at N328 (of the NODAL variant) were both predicted to be N-glycosylated (Figure 4.4 and 4.5).

Next, cells were subject to different treatments to determine the nature of the different bands. First, cyclohexamide was used to arrest translation to assess the dynamics of each NODAL peptide species. After 24 hours of treatment, the smaller NODAL peptides had decreased in intensity as expected in the absence of de novo translation, while the largest peptide for each species actually accumulated and increased in intensity after 24 hours of treatment (Figure 4.6A). This suggested that the difference between the bands was the result of a post-translational process. Next, tunicamycin was used to block global N-glycosylation of proteins, which resulted in partial or complete loss of all NODAL bands for both proteoforms, and the emergence of a smaller band (Figure 4.6B). Finally, mutation of N72 and N199 in the constitutive NODAL protein recapitulated the tunicamycin result, while the largest band was lost upon mutation of N328 in the NODAL variant protein (Figure 4.6C). Collectively, these results suggest that the NODAL proteins are differentially N-glycosylated. One proteoform is likely modified at one site and unmodified at the other, while the other proteoform is likely modified at both

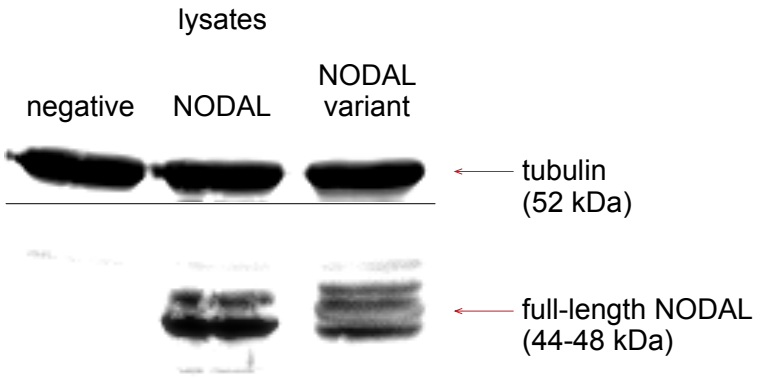


Figure 4.3: Stable expression of both NODAL isoforms reveals multiple bands in HEK 293 cell lysates.
Approximate sizes of detected bands are indicated. Constitutive NODAL revealed two bands. NODAL variant revealed three bands. Tubulin was included as a loading control, and NODAL was detected with an anti-Myc tag antibody.

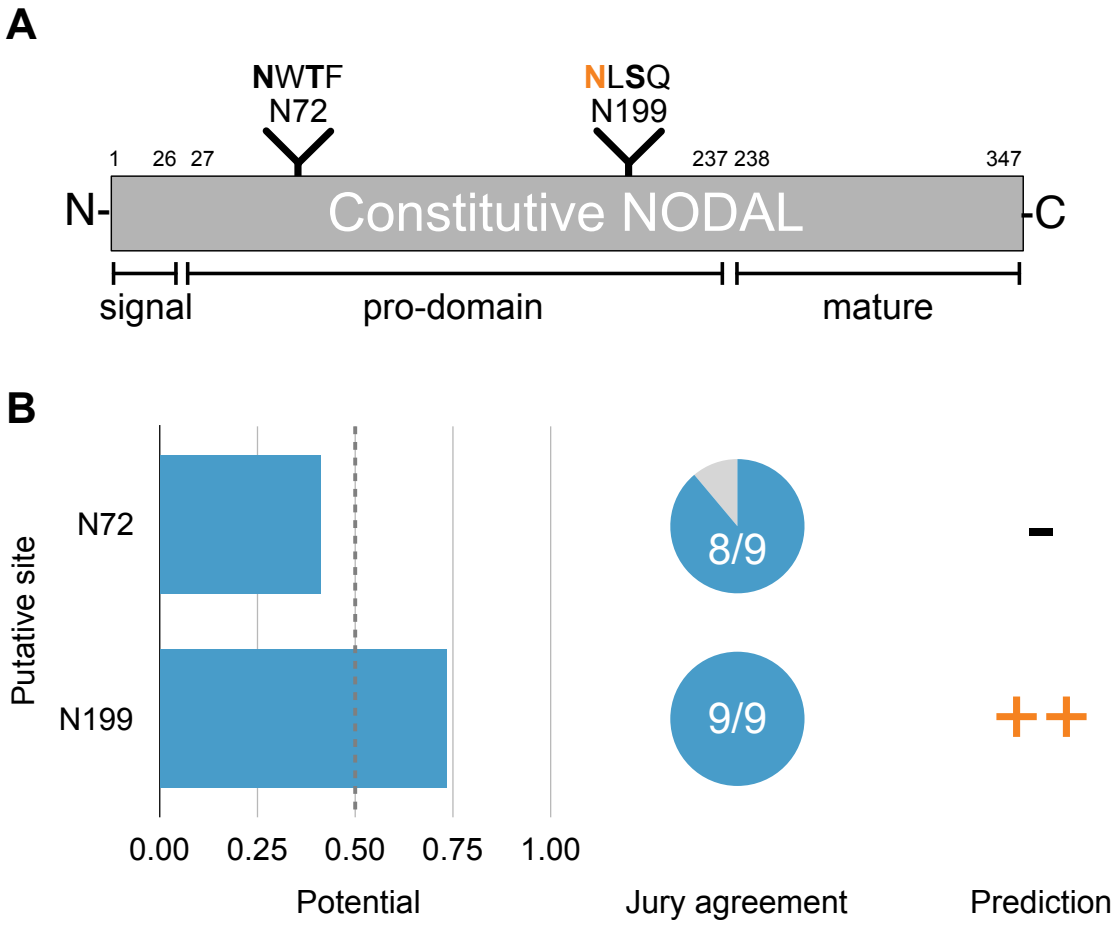


Figure 4.4: NODAL is predicted to be N-glycosylated at one of two N-glycosylation motifs in the pro-domain.

A) Constitutive NODAL protein is illustrated with the N-terminus on the left and C-terminus on the right. Numbers mark amino acid positions for the start and end of each peptide element. “Y” indicates positions of potentially N-glycosylated asparagine residues. Sequences above asparagine residues show the complete NX[ST]X (where “X” is any amino acid except for proline) N-gylcosylation motif context for each site. “N” in black indicates a site predicted to remain unmodified. “N” in orange indicates a site predicted to be N-glycosylated. B) N-glycosylation prediction for each motif from the NetNGlyc 1.0 Server. Dashed grey line at potential = 0.5 indicates general threshold for positive N-glycosylation prediction. “-” prediction result indicates threshold < 0.5. “++” indicates potential >0.5 and jury agreement 9/9. See methods for full range of possible predictions.

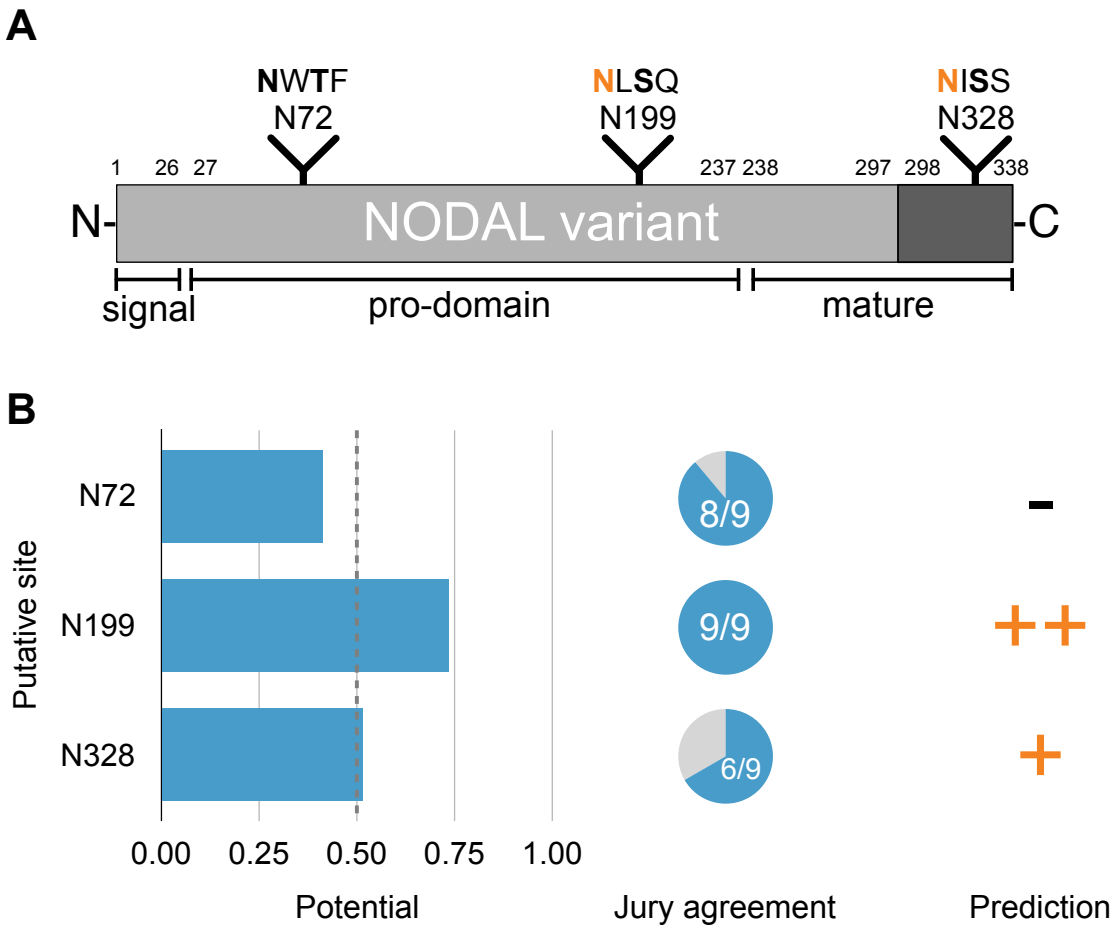


Figure 4.5: The NODAL variant is predicted to be N-glycosylated at a novel N-glycosylation motif in the mature peptide.

A) NODAL variant protein is illustrated with the N-terminus on the left and C-terminus on the right. Numbers mark amino acid positions for the start and end of each peptide element. “Y” indicates positions of potentially N-glycosylated asparagine residues. Sequences above asparagine residues show the complete NX[ST]X (where “X” is any amino acid except for proline) N-gylcosylation motif context for each site. “N” in black indicates a site predicted to remain unmodified. “N” in orange indicates a site predicted to be N-glycosylated. B) N-glycosylation prediction for each motif from the NetNGlyc 1.0 Server. Dashed grey line at potential = 0.5 indicates general threshold for positive N-glycosylation prediction. “-” prediction result indicates threshold < 0.5. “+” indicates potential > 0.5. “++” indicates potential >0.5 and jury agreement 9/9. See methods for full range of possible predictions.

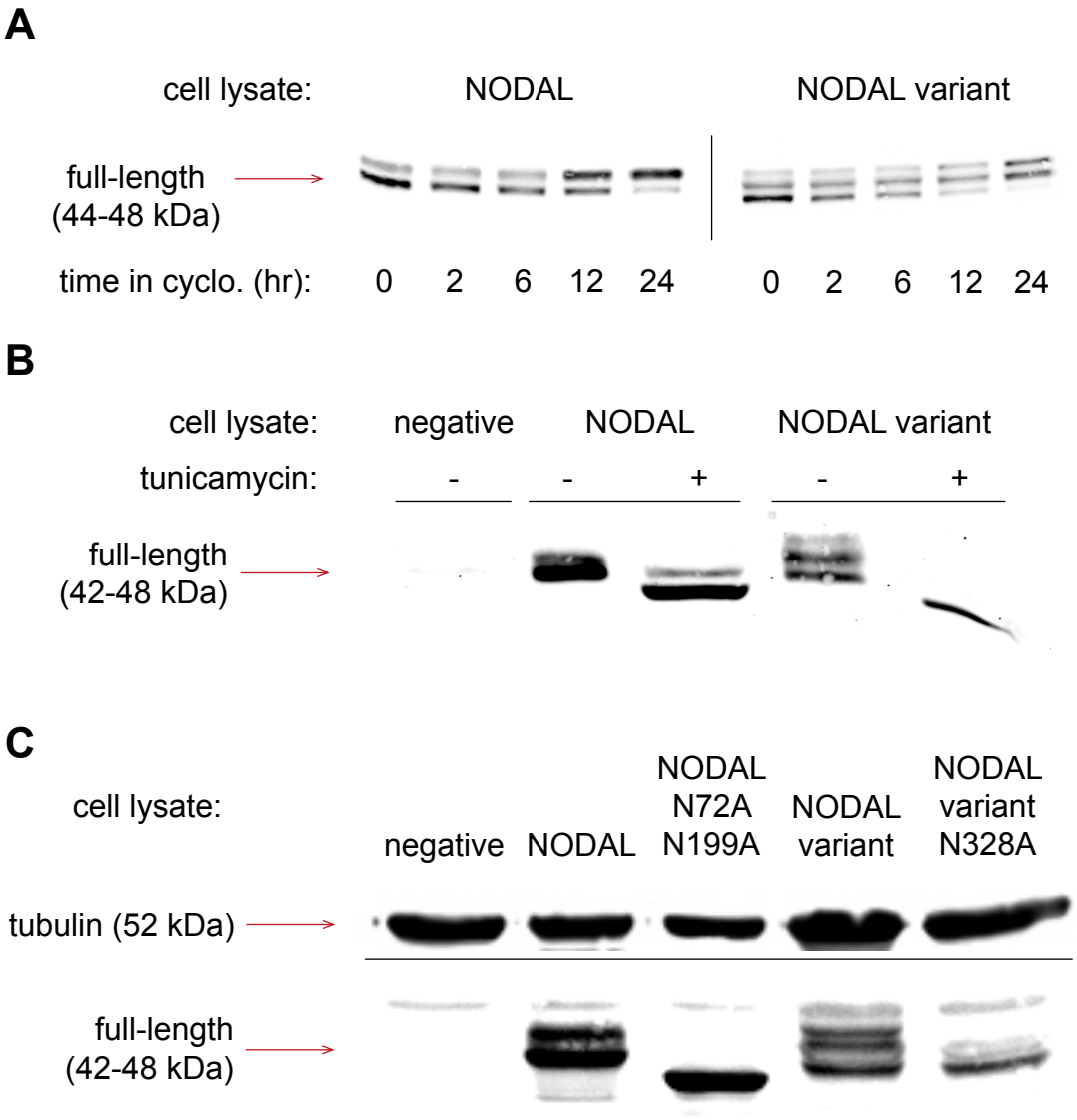


Figure 4.6: Both full-length NODAL isoforms display distinct patterns of N-glycosylation.
NODAL is differentially N-glycosylated at two sites, while the NODAL variant is differentially N-glycosylated at three sites. A) Western blot of NODAL or NODAL variant-expressing cells treated with cyclohexamide. “cyclo.” = cyclohexamide. B) Western blot of NODAL or NODAL variant-expressing cells treated with tunicamycin. C) Western blot of NODAL and NODAL variant N-glycosylation motif mutants. Tubulin was included as a loading control. NODAL was detected with an anti-Myc tag antibody. Approximate sizes of detected bands are shown. An equal amount of protein was loaded for each sample. A) and B) Amido black staining of membranes was used to verify equal protein transfer.

sites. Similarly, the NODAL variant is likely modified at either one, two, or all three sites.

Since N-glycosylation is a common modification for secreted proteins, I was interested characterizing the secretion of each NODAL proteoform and assess if the novel N-glycosylation had any impact on secreted protein. Collection of serum-free conditioned media from over-expressing cells revealed both full-length and processed NODAL peptides for both NODAL and the NODAL variant (Figure 4.7). Secreted NODAL variant protein with a mutated N-glycosylation site in the mature peptide also revealed a shift in the size of the mature processed peptide, confirming alternative N-glycosylation of this site. Furthermore, the mature NODAL peptides had different profiles, again indicative of differential post-translational modification. The constitutive NODAL isoform had two bands with a very small difference in size. Mature NODAL variant had a similar profile, with two sets of two bands each, for each of the N-glycosylation states (Figure 4.7). Since this modification is shared between both isoforms, it is therefore likely to take place in the N-terminal half of the mature domain.

The ratio of mature:full-length NODAL was determined using the integrated intensities of bands detected in the conditioned media. This ratio did not differ between NODAL, the NODAL variant, or the NODAL variant with a mutated N-glycosylation motif at N328 according to an ANOVA test ($P = 0.340$; Figure 4.8A). However, the ratio of total NODAL protein in the media relative to the lysate was higher for the NODAL variant than the constitutive NODAL proteoform, and this difference was partially restored to constitutive NODAL levels upon mutation of N328, according to an ANOVA test ($P = 0.041$; Figure 4.8B). These results suggest that the NODAL variant is either more efficiently secreted or stabilized in the media relative to constitutive NODAL, and that the former's unique N-glycosylation is at least partially responsible for this effect.

To determine whether preferential accumulation of NODAL variant could explain its increased extracellular presence, conditioned media collected from NODAL-expressing cells was transferred to naive untransfected cells where no newly translated and secreted NODAL would interfere with analysis (Figure 4.9A). This system allowed tracking of

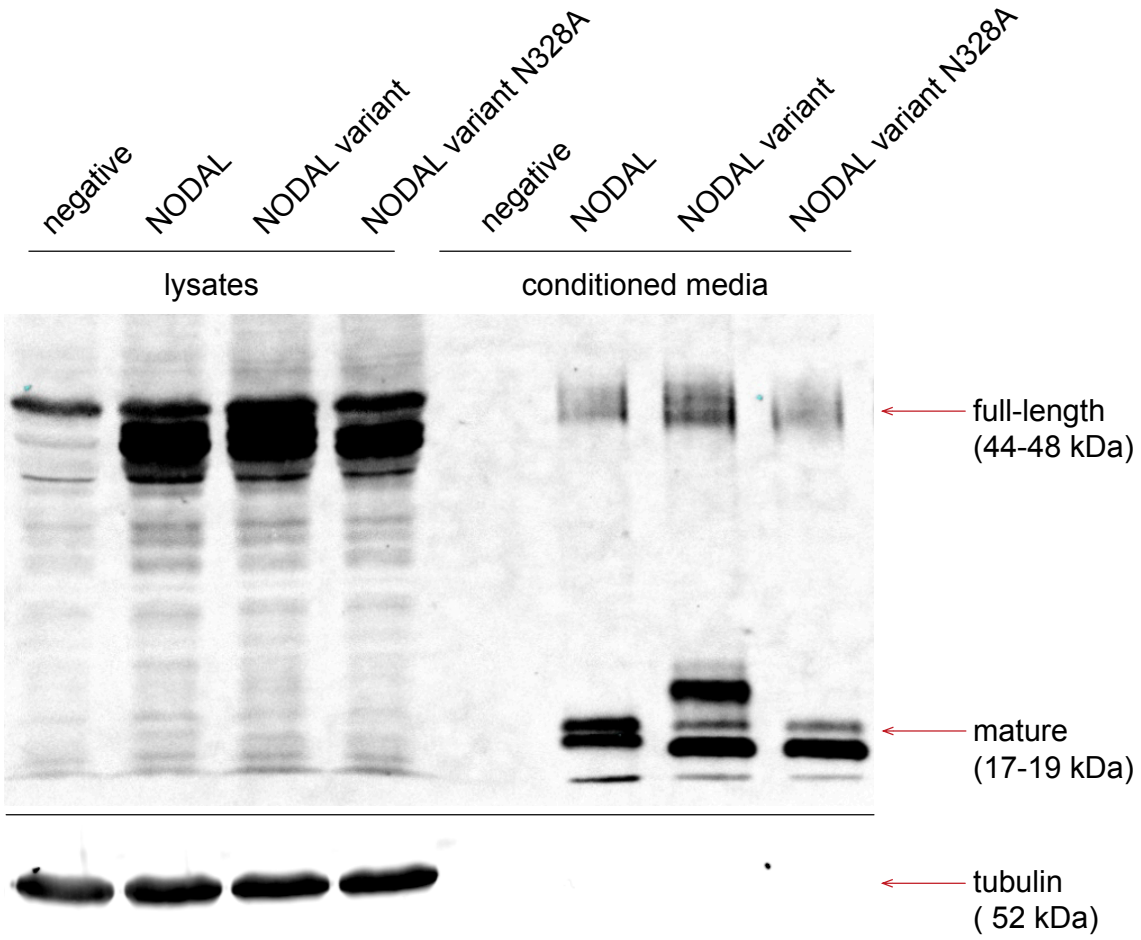


Figure 4.7: Both NODAL proteoforms are present in conditioned media. Both NODAL isoforms are processed and mature peptides for each isoform are differentially post-translationally modified in distinct fashions. Mutation of NODAL variant N328 results in loss of a larger mature NODAL variant band and a banding pattern that more closely resembles that of constitutive NODAL. Only full-length NODAL peptides are present in corresponding cell lysates. Approximate sizes of detected bands are shown. Cell lysate from the same number of cells, and conditioned media from the same number of cells were analyzed for each sample. Image light was adjusted for clear visualization of full-length NODAL peptides in conditioned media, resulting in less clear definition of NODAL bands in corresponding cell lysates. Tubulin was included as a loading control, and NODAL was detected with an anti-Myc tag antibody. A representative image from two analyses is shown.

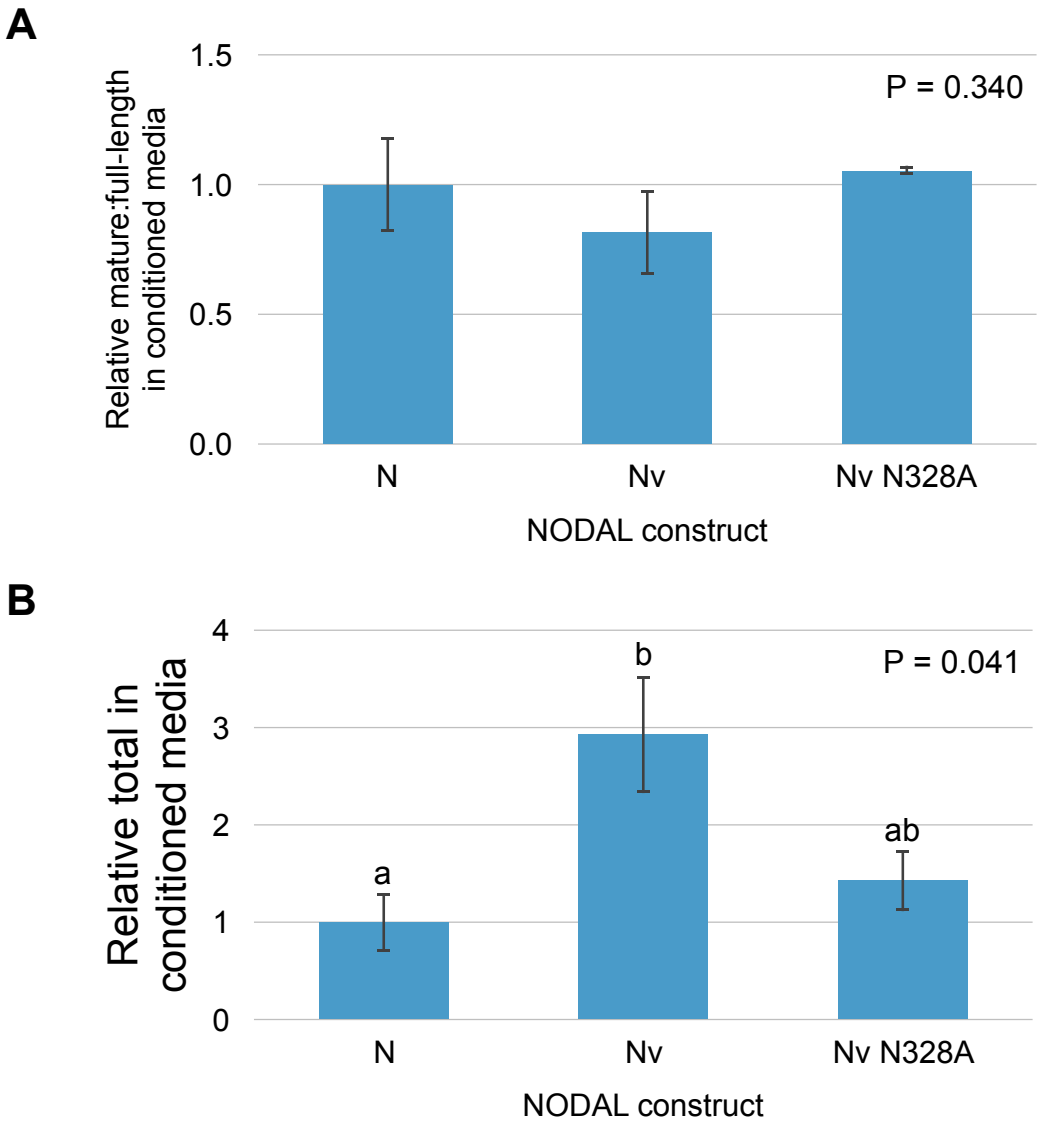


Figure 4.8: The NODAL variant protein is preferentially secreted relative to constitutive NODAL.

This difference is reduced upon abrogation of N-glycosylation of the mature peptide of the NODAL variant. “N” = constitutive NODAL. “Nv” = NODAL variant. Error bars indicate standard deviations. P values shown are results of ANOVA tests for all three time constructs. For ANOVA tests with $P < 0.05$, Tukey HSD post hoc tests were performed. Different letters (e.g. ‘a’ and ‘b’) indicate a statistically significant ($P < 0.05$) difference between two samples according to the Tukey HSD test. Pairs of samples with the same letter (e.g. ‘a’), are not statistically different.

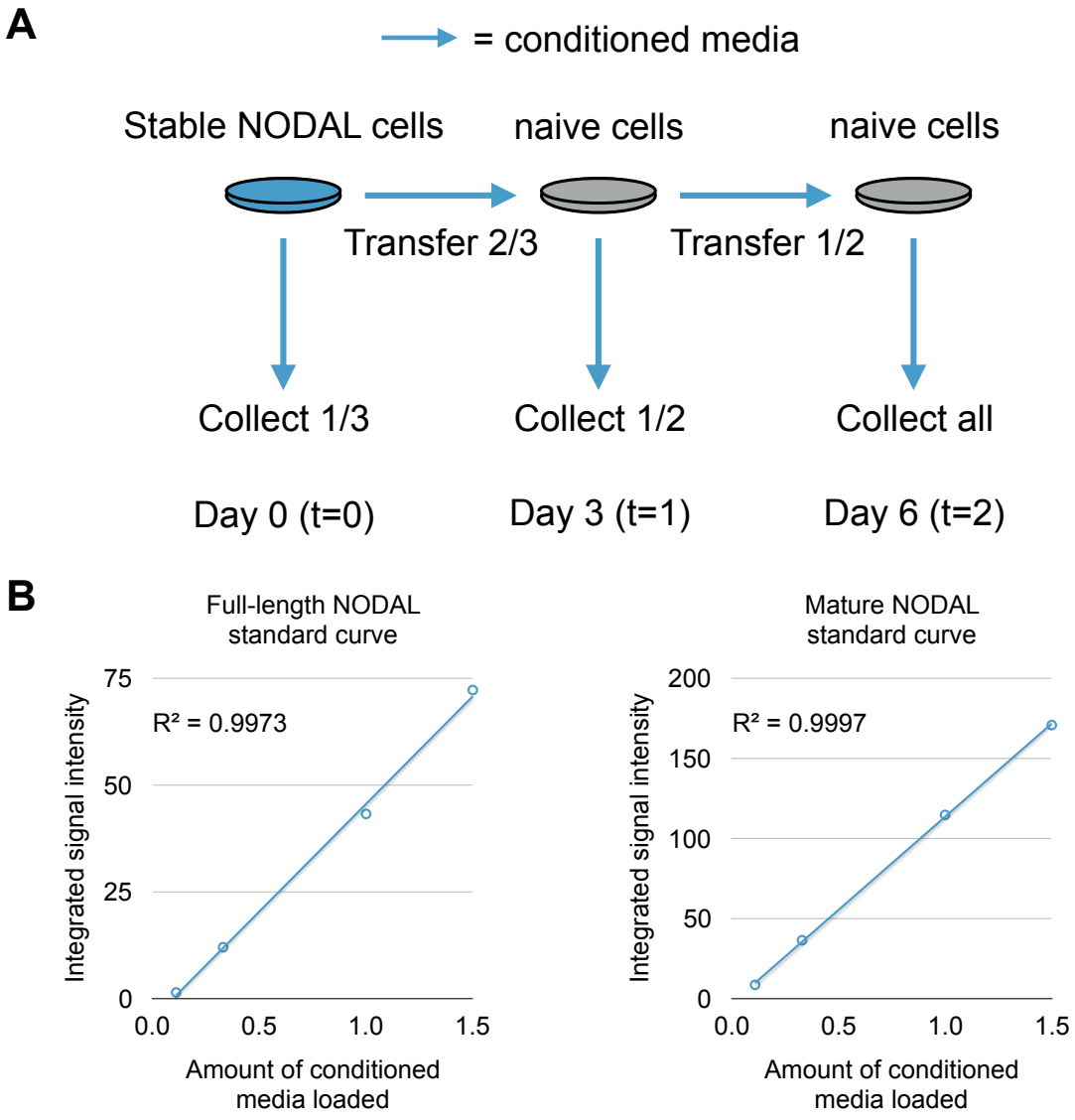


Figure 4.9: Development of a conditioned media transfer system.
This system was used to quantitatively study NODAL protein processing and breakdown in the absence of chemical inhibitors. A) Schematic of methodology used. Identical volumes of the original conditioned media were collected at each time point. B) Validation of the quantifiable linear range of western blot assays used to quantify NODAL levels in conditioned media. R^2 values indicate the coefficient of determination for full-length and mature peptides. Examples shown are for NODAL. Similar standard curves were utilized for each construct tested.

natural NODAL processing (from full-length to mature peptide), and protein break-down/ turnover. Standard curves corresponding to each sample were used to ensure accurate quantification (Figure 4.9B). As expected, constitutive NODAL was continually processed in the media (Figure 4.10A), resulting in less full-length protein over time (Figure 4.10B). Mature NODAL protein remained unchanged after three days, but began to decrease by day six (Figure 4.10C). Consequently, total NODAL levels remained unchanged after three days, and began to decrease after six days, while the ratio of mature:full-length NODAL increased over time (Figure 4.10D). These experiments revealed similar dynamics between constitutive NODAL, the NODAL variant, and the novel N-glycosylation-mutated NODAL variant (Figure 4.11). The levels of total NODAL protein in the media did not differ between constitutive NODAL and NODAL variant after either three or six days (Figure 4.11C). This suggests that increased secretion, and not increased intrinsic stability, is responsible for increased NODAL variant in the media. Interestingly, the accumulation of mature protein relative to its full-length precursor was more prominent for the NODAL variant relative to constitutive NODAL (Figure 4.11D). The ratio of mature:full-length protein was partially restored to constitutive NODAL levels in the NODAL variant N-glycosylation mutant, suggesting N-glycosylation of the NODAL variant may confer a small stabilizing effect on the mature peptide.

These findings for the NODAL variant led me to test whether N-glycosylations in the NODAL pro-domain common to both NODAL proteoforms also impact the amount of NODAL protein in conditioned media. Dual mutation of N72 and N199 residues resulted in a decrease in the amount of total NODAL present in the conditioned media relative to the cell lysate according to a t-test ($P = 0.009$; Figure 4.12A,C). Interestingly, the ratio of mature:full-length NODAL in the conditioned media was also increased upon loss of N-glycosylation in the pro-domain according to a t-test ($P = 0.075$; Figure 4.12A,B). In a conditioned media transfer experiment, dual mutation of N-glycosylation motifs did not reduce full-length or total NODAL levels relative to unmutated protein after three or six days (Figure 4.13). Consequently, there was also no corresponding increase in the mature:full-length NODAL ratio in the N-glycosylation mutant after three or six days (Figure 4.13). Collectively, these results suggest N- glycosylations promote secretion of

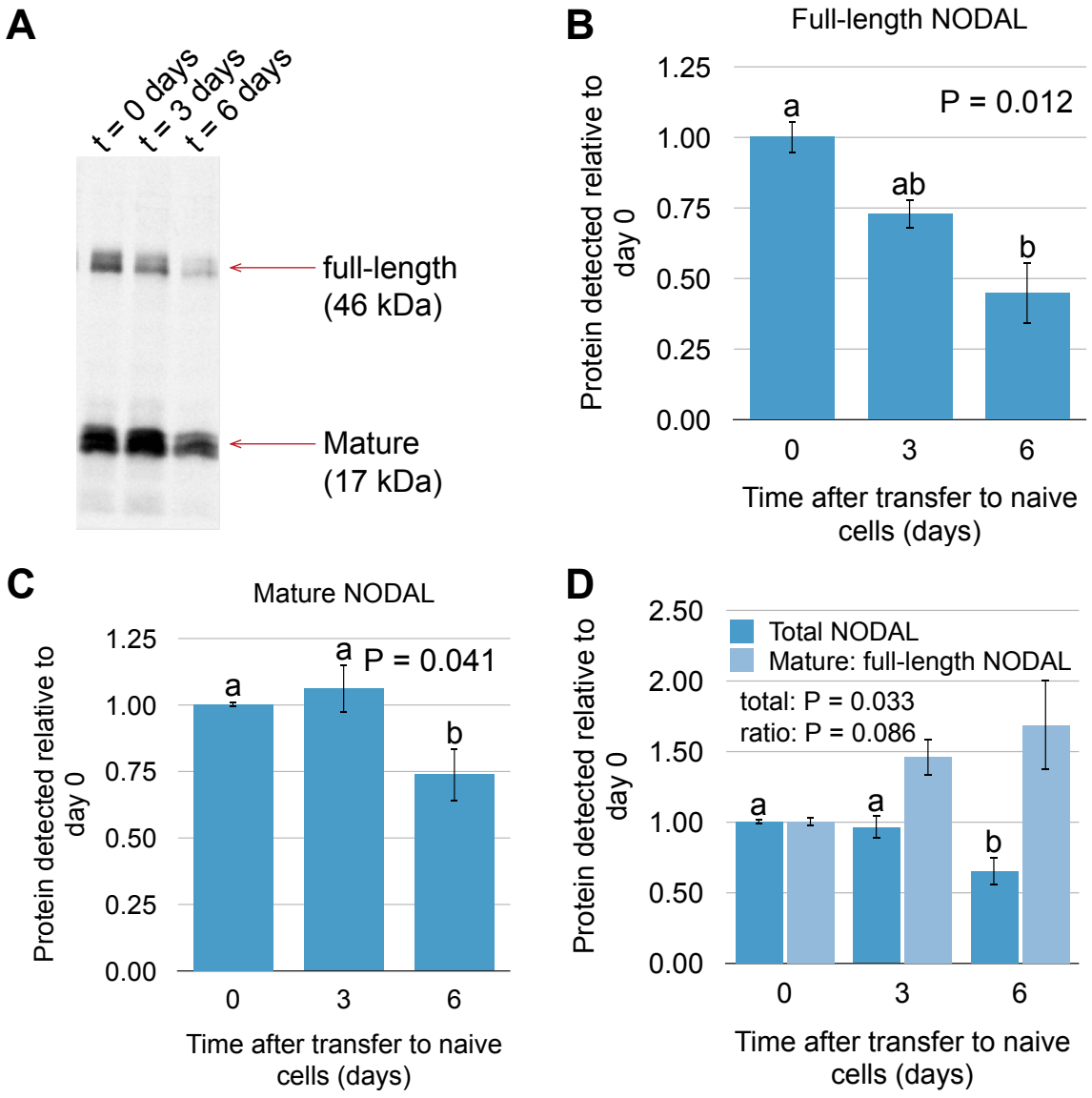


Figure 4.10: All of full-length, mature, and total constitutive NODAL protein shows significant reduction after six days in protein turn-over experiments. A) A representative Western blot of constitutive NODAL processing and break down over time in conditioned media. NODAL was detected with an anti-Myc tag antibody. A representative image from two analyses is shown. Approximate sizes of detected bands are shown. B-D) Quantification of constitutive NODAL peptides in cell culture media. Error bars indicate standard deviations. P values shown are results of ANOVA tests for all three time points. For ANOVA tests with $P < 0.05$, Tukey HSD post hoc tests were performed. Different letters (e.g. 'a' and 'b') indicate a statistically significant ($P < 0.05$) difference between two samples according to the Tukey HSD test. Pairs of samples with the same letter (e.g. 'a'), are not statistically different.

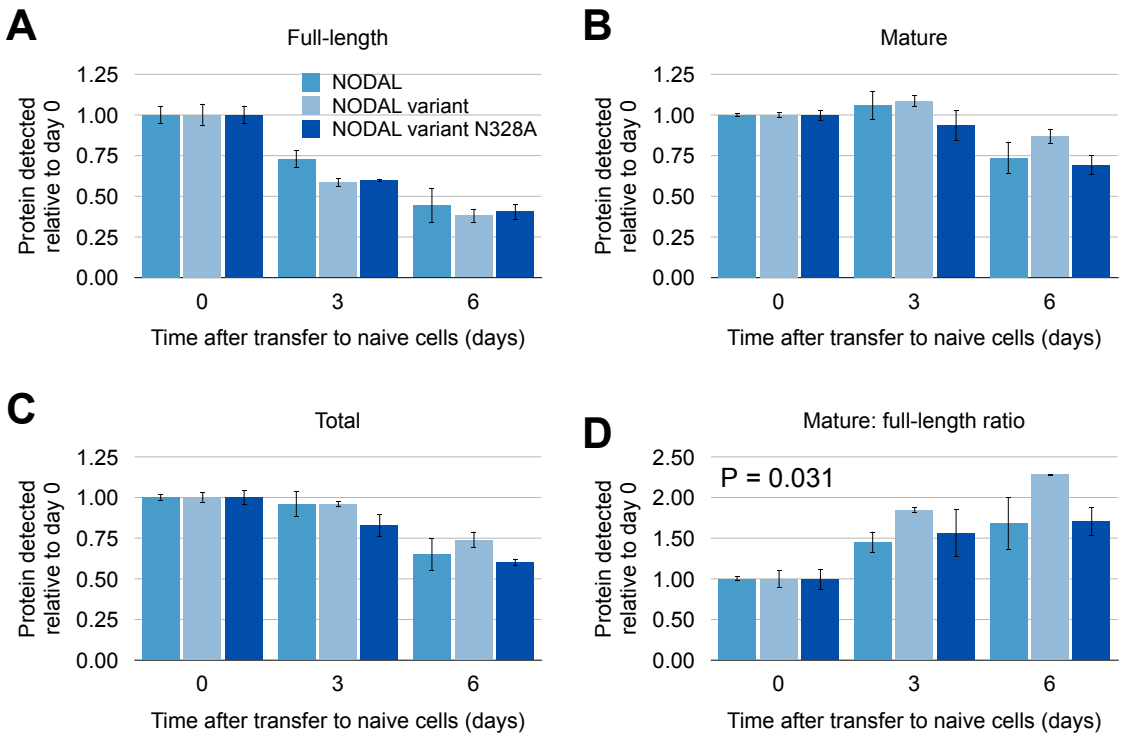


Figure 4.11: The NODAL variant displays a small increase in the mature: full-length peptide ratio relative to constitutive NODAL.

This difference is diminished when N-glycosylation of the mature NODAL variant peptide is inhibited. Error bars indicate standard deviations. P value in D is the significance test result for differences between NODAL constructs according to ANCOVA at both three and six days after conditioned media transfer. None of full-length, mature, or total protein showed significant differences across both time points (all $P > 0.05$ by ANCOVA).

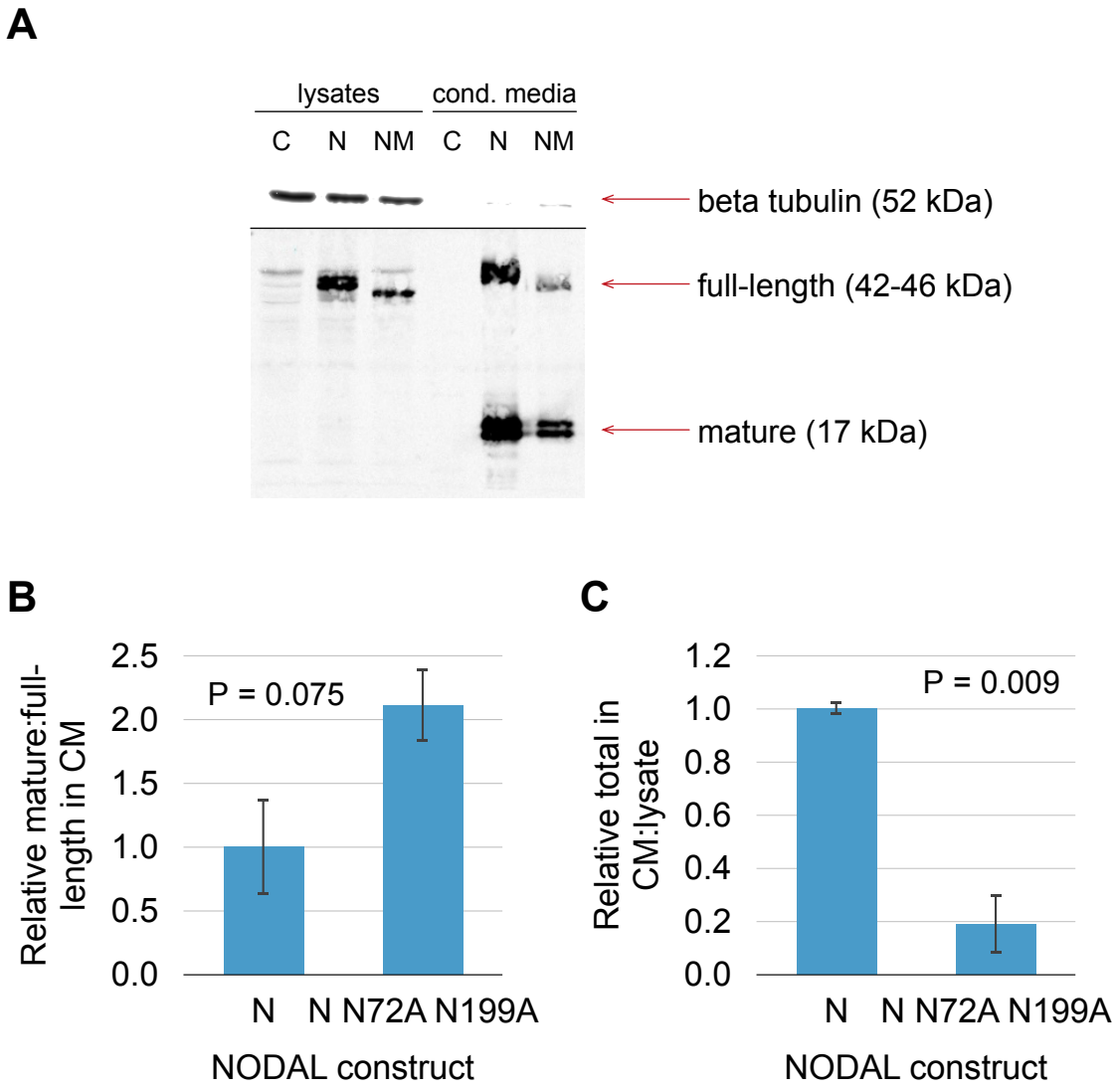


Figure 4.12: N-glycosylation of the NODAL pro-domain affects NODAL processing.

Mutation of both N-glycosylation motifs in the NODAL pro-domain results in dramatically reduced conditioned media:cell-lysate ratios of NODAL protein and an increase in the mature:full-length ratio in conditioned media. Cell lysate from the same number of cells, and conditioned media from the same number of cells were analyzed for each sample. A) "C" = control. "N" = NODAL. "NM" = NODAL N72A N199A double mutant. Approximate sizes of detected bands are shown. B-C) P values are results of significance test for differences between NODAL and dual N-glycosylation mutant NODAL by t-test. Tubulin was included as a loading control. NODAL was detected with an anti-Myc tag antibody. A representative image from two analyses is shown.

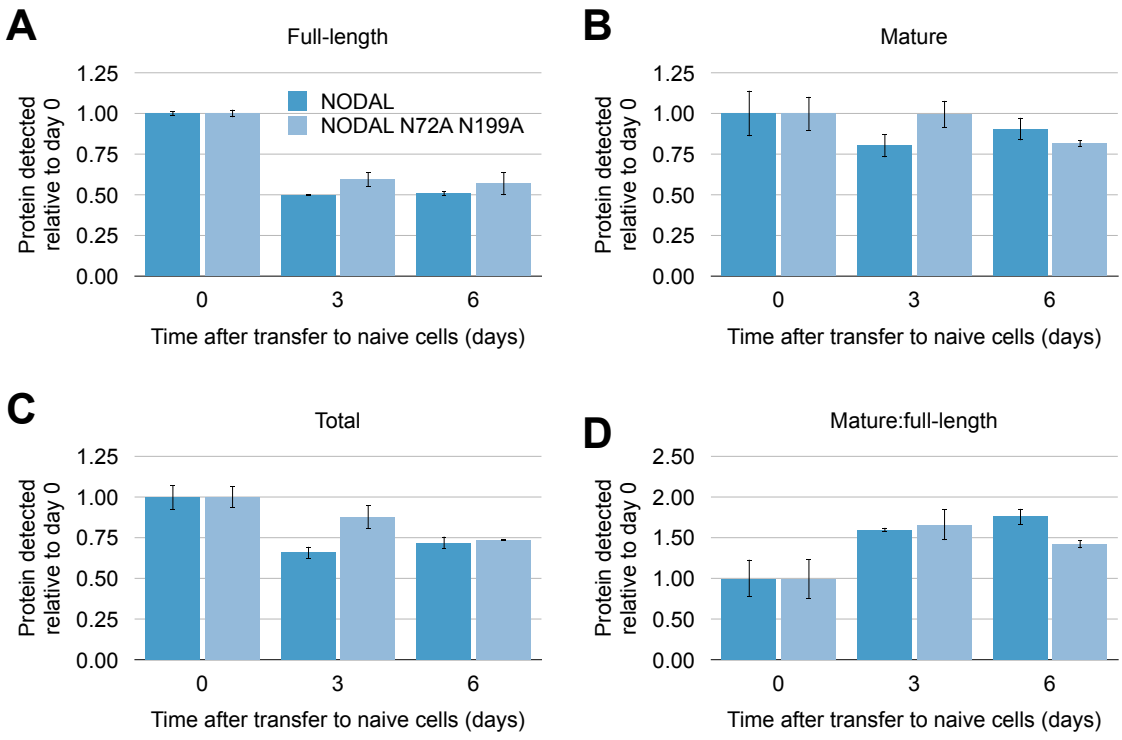


Figure 4.13: Loss of NODAL N-glycosylations has no consistent effect on protein break-down in conditioned media.

Error bars indicate standard deviations. None of full-length, mature, total, or mature:full-length protein showed significant differences across both time points (all $P > 0.05$ by ANCOVA).

NODAL protein, and may regulate the processing of full-length NODAL, but do not stabilize extracellular NODAL protein *in vitro*.

Next, I was interested in comparing the overall structure of the mature peptides for each NODAL proteoform. A conserved protein domain search for the NODAL variant mature peptide sequence revealed a partial TGF-beta family domain. Interestingly, several amino acids downstream of exon 2 and unique to the NODAL variant contributed to a TGF-beta family domain signature (Figure 4.14). Secondary structure prediction using JPred predicted similar stretches of secondary structure between the two NODAL isoforms, even for the novel C-terminus of the NODAL variant (Figure 4.15). Notable differences include a truncated alpha-helix towards the middle of the NODAL variant protein, and three rather than four segments of beta-sheet in the unique C-terminal half of the protein.

One important element of TGF-beta superfamily members is a conserved set of six cysteine residues that form an intricate structure known as a cysteine knot [41]. Constitutive human NODAL contains seven cysteines in its mature peptide. Disulfide bond prediction analysis predicted disulfide bonds between cysteines 1 and 5, 2 and 6, and 3 and 7 characteristic of a cysteine knot (Figure 4.16A and B). Intriguingly, the NODAL variant mature peptide also contains exactly seven cysteines, with very similar spacing relative to those found in NODAL, despite four of these cysteines being coded downstream of the shared constitutive exon 2. These cysteines, however, are not predicted to form disulfide bonds in a pattern resembling a TGF-beta-like cysteine knot (Figure 4.16C).

As an extension of these secondary structure and disulfide bond predictions, I compared predicted protein structures of NODAL and NODAL variant mature proteins. While no crystal structure has been reported for the mature NODAL peptide itself, there is a crystal structure of the NODAL:BMP2 chimeric protein with NODAL function [24] introduced above. This protein contains a large segment of NODAL sequence, with many other shared and similar residues to NODAL (Figure 4.17). A predicted structural model for the NODAL mature peptide reveals a very similar structure to the NODAL:BMP2 chimera (Figure 4.18). A predicted structure for the NODAL variant was also generated.

Figure 4.14: Partial conservation of a TGF-beta family domain in the mature NODAL variant peptide.

Red amino acids indicate exact matches between NODAL and the TGF-beta family domain “smart00204.” All other amino acids are blue. “-” indicates a gap in the alignment. Numbers above sequence indicate relative position in mature peptide from N-terminus (1) toward the C-terminus.

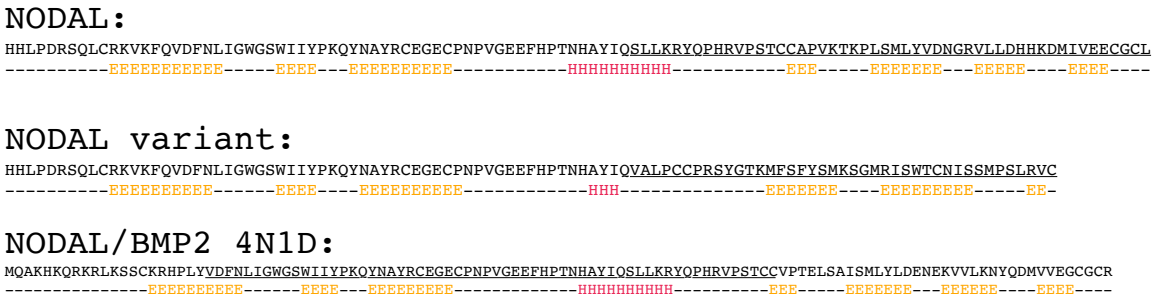


Figure 4.15: Similar secondary structures are predicted for the mature peptides of NODAL and the NODAL variant.
Underlined amino acids in NODAL sequences indicate residues unique to NODAL or the NODAL variant coded for downstream of constitutive exon 2. Underlined amino acids for the NODAL/BMP2 chimera mark the segment with an exact match to NODAL. “E” indicates residues predicted to adopt a beta-sheet secondary structure. “H” indicates residues predicted to adopt a helical structure.

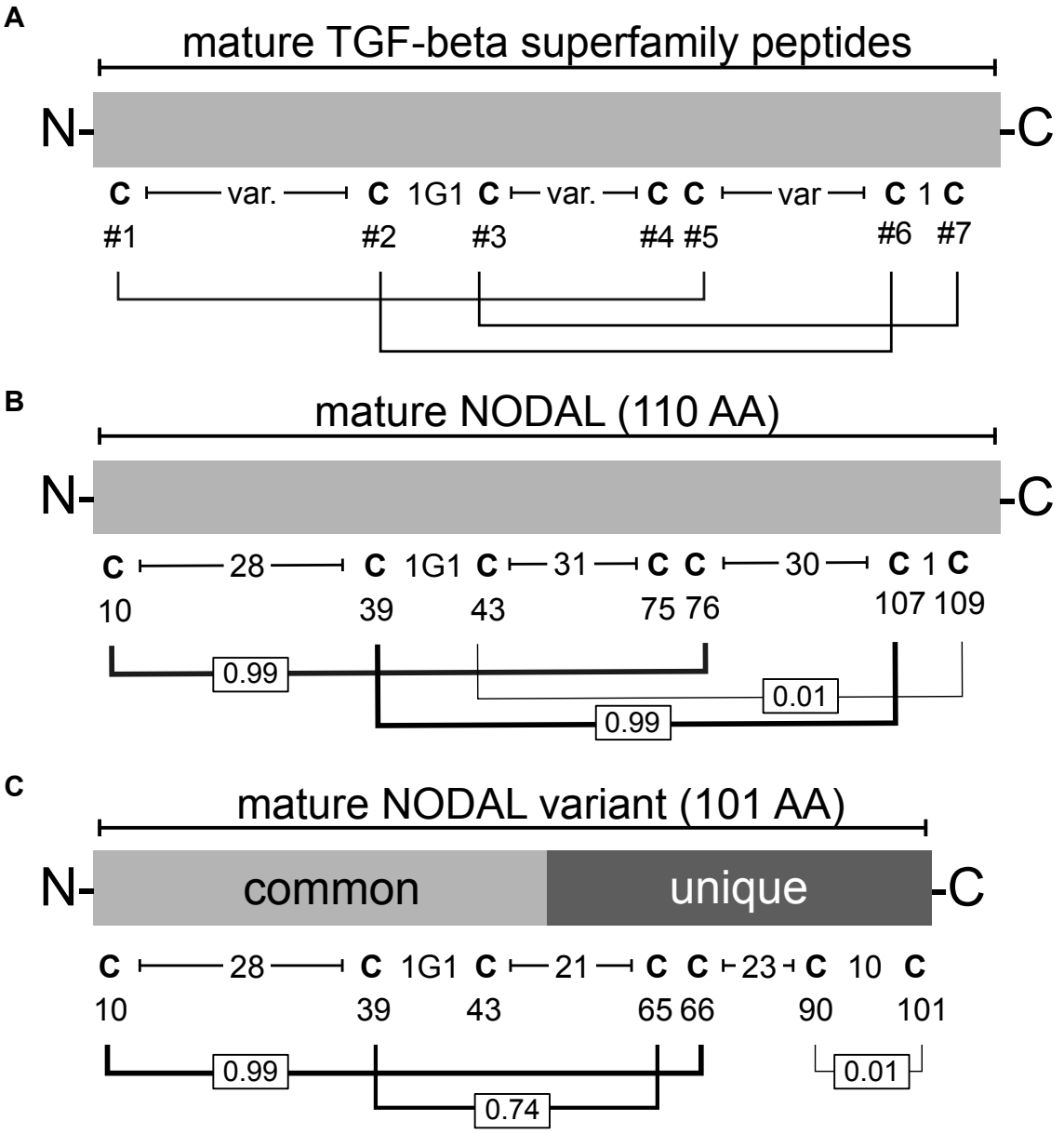


Figure 4.16: The NODAL isoforms are predicted to form different intrachain disulfide bonds, with only constitutive NODAL forming bonds characteristic of a cysteine knot.

A) “var.” = variable number of amino acids between flanking cysteines. Numbers in first row below protein schematics in B) and C) indicate number of residues between adjacent cysteines. Numbers below cysteines indicate their position along the mature peptide from N-terminus to C-terminus. Lines connecting cysteines indicate predicted disulfide bonds, with their thickness positively correlated to the score of the predicted bond. Predicted bond score (ranging from 0-1) is indicated within a box on each line.

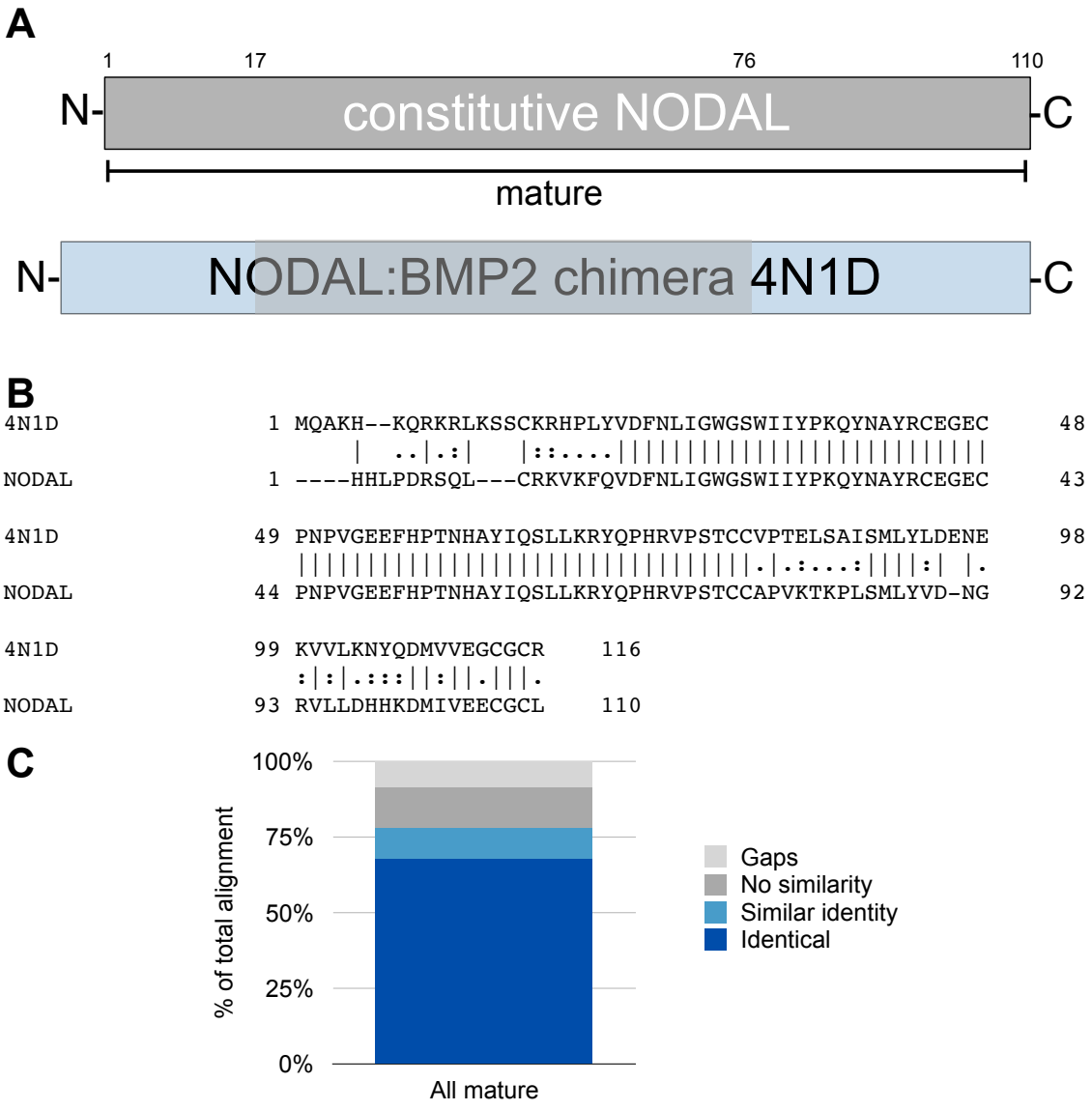


Figure 4.17: A NODAL:BMP2 chimera with known structure is similar in amino acid identity to human NODAL.

A) Numbers above protein schematic indicate start (17) and end (76) positions of continuous NODAL sequence in the NODAL:BMP2 chimera 4N1D, also indicated by grey shading of the chimeric schematic. B) "|" = exact match amino acid pairs. ":" = similar amino acids. "." = no similarity. "-" = gap in alignment. C) Aligned amino acids by category.

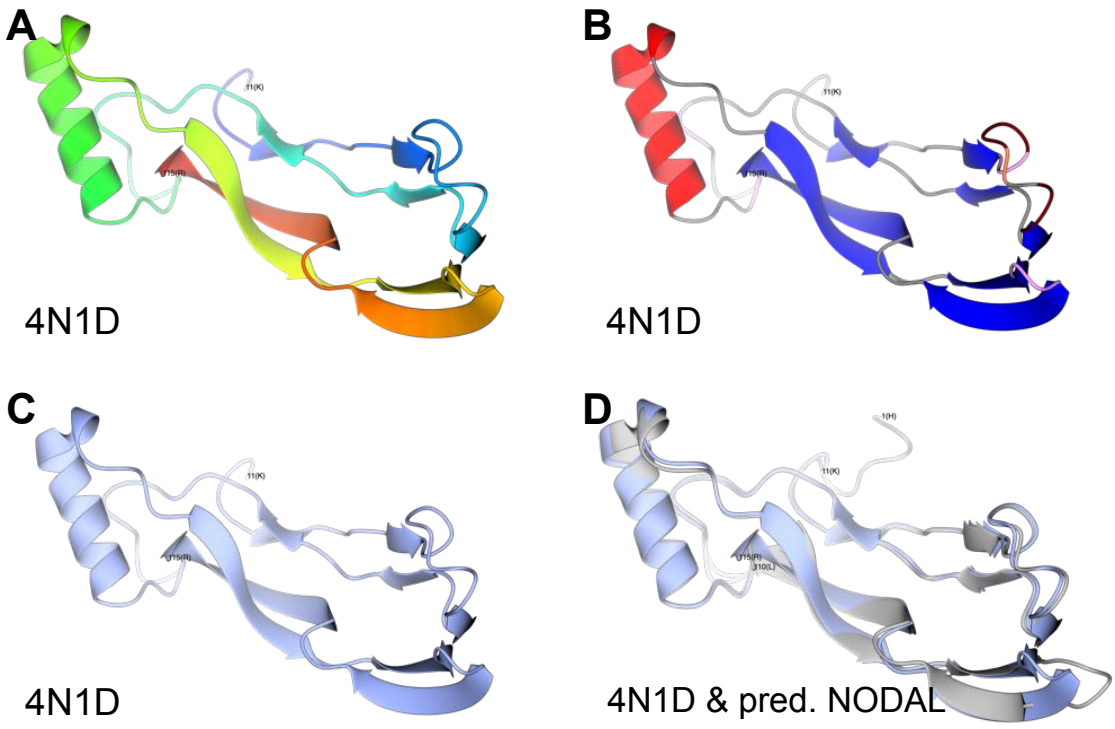


Figure 4.18: NODAL is predicted to have a similar structure to NODAL:BMP2 chimera NB250 (4N1D).
Superimposition of a predicted structure for human NODAL and 4N1D revealed a high degree of similarity. First (N-terminal) and last (C-terminal) amino acids are labelled for each structure. A) Rainbow of 4N1D chimera structure, from N-terminus (violet) to C-terminus (red). B) 4N1D chimera structure with secondary structure shown. Blue = beta strand/ beta bulge, grey = no structure, pink = 3-turn, tan = 4-turn, coral = 5-turn, red = alpha helix. C) 4N1D chimera without colour-coding. D) Superimposition of 4N1D (blue) and a structure predicted for the constitutive NODAL mature peptide (grey) by Phyre2.

This structure differed from the chimera structure in that the “wrist” alpha-helical structure is truncated, and the C-terminal end of the protein does not extend all the way back to the cysteine knot structure. The NODAL variant is predicted to contain two C-terminal anti-parallel beta sheet structures that form two “finger” projections similar to constitutive NODAL and the chimera structure (Figure 4.19). Additionally, while half of the ring structure of the cysteine knot is absent in the NODAL variant, there is a disulfide bond that passes through the half ring area in an identical fashion to constitutive NODAL (Figure 4.20).

Aside from the six cysteine residues involved in intrachain disulfide bond formation, there is a seventh cysteine at position 312 of constitutive NODAL putatively involved in NODAL:NODAL homo-dimerization through the formation of an interchain disulfide bond. This function is inferred by similarity, and has never been directly experimentally studied for human NODAL. I sought to investigate the role of C312 on NODAL protein dynamics. Expression of NODAL with C312S mutation resulted in both decreased total NODAL detected in conditioned media relative to cell lysates, and a large increase in the mature:full-length peptide ratio in the conditioned media (Figure 4.21). Notably, C312S mutation also resulted in a higher molecular weight mature peptide, perhaps indicative of cryptic post-translational modification. In conditioned media transfer experiments, there was no consistent effect on protein processing and turnover dynamics upon loss of C312 (Figure 4.22).

I was also interested in assessing the dimerization capacity of NODAL proteoforms and different NODAL mutants in the media. Non-reducing SDS PAGE and subsequent Western blot analysis was employed to specifically detect size-shifted NODAL complexes indicative of interchain disulfide bond formation. For cell lysates, NODAL did not reveal any discrete complex formation, as only full-length NODAL protein was clearly evident. However, both the NODAL variant and NODAL variant N328A proteins revealed at least one discrete complex (Figure 4.23A). In corresponding conditioned media, bands consistent with only full-length and mature NODAL peptides were evident for NODAL. However, the NODAL variant and NODAL variant N328A revealed at least

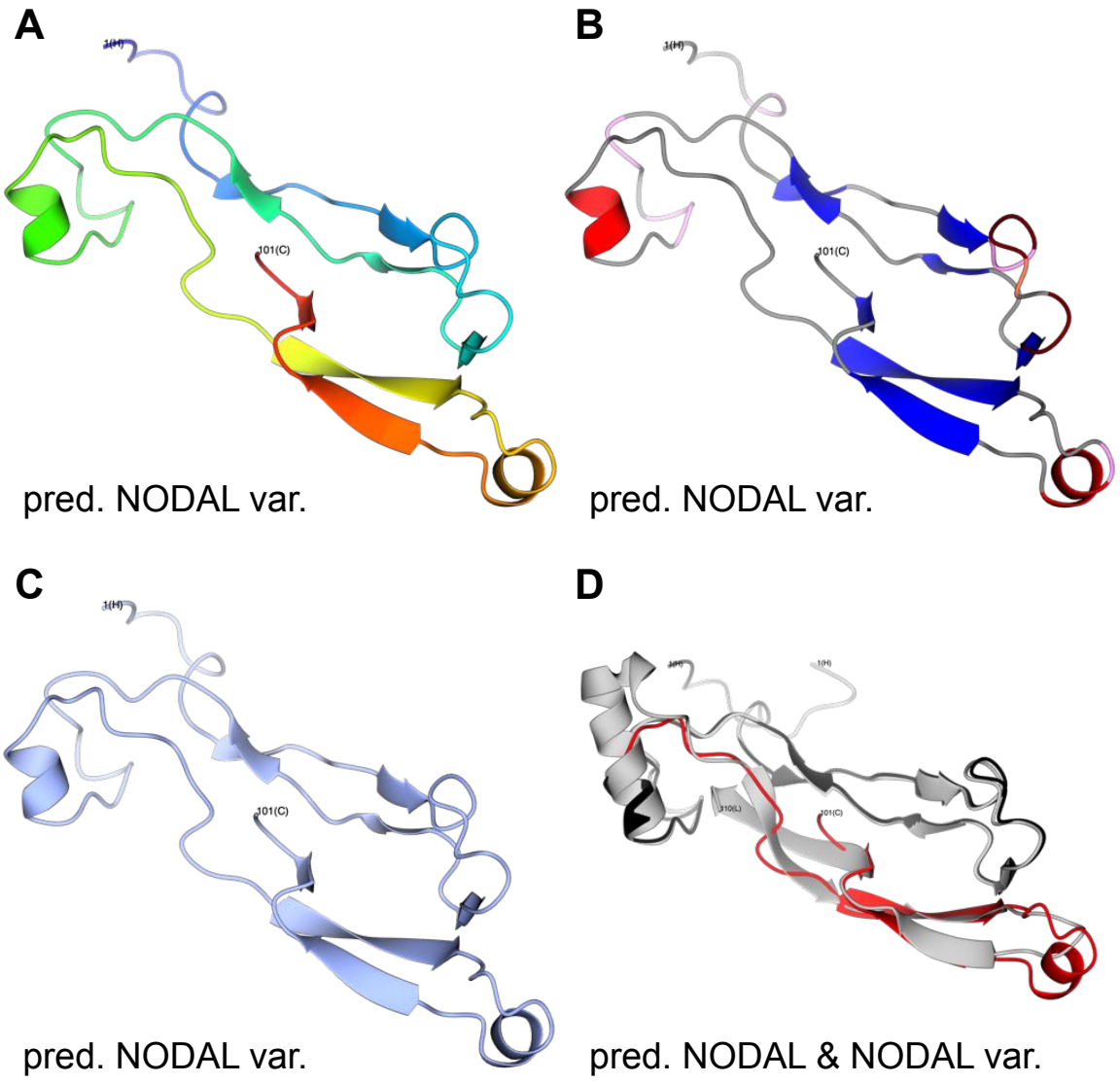


Figure 4.19: A predicted structure for the human NODAL variant is distinct from the experimentally determined structure for NODAL:BMP2 chimera NB250 (4N1D).

First (N-terminal) and last (C-terminal) amino acids are labelled for each structure. A) Rainbow of mature NODAL variant predicted structure, from N-terminus (violet) to C-terminus (red). B) mature NODAL variant predicted structure with secondary structure shown. Blue = beta strand/ beta bulge, grey = no structure, pink = 3-turn, tan = 4-turn, coral = 5-turn, red = alpha helix. C) mature NODAL variant predicted structure without colour-coding. D) Superimposition of a structure predicted for the mature NODAL variant peptide (dark grey is sequence common to NODAL, red is unique sequence coded by downstream of constitutive exon 2), and predicted NODAL structure (light grey).

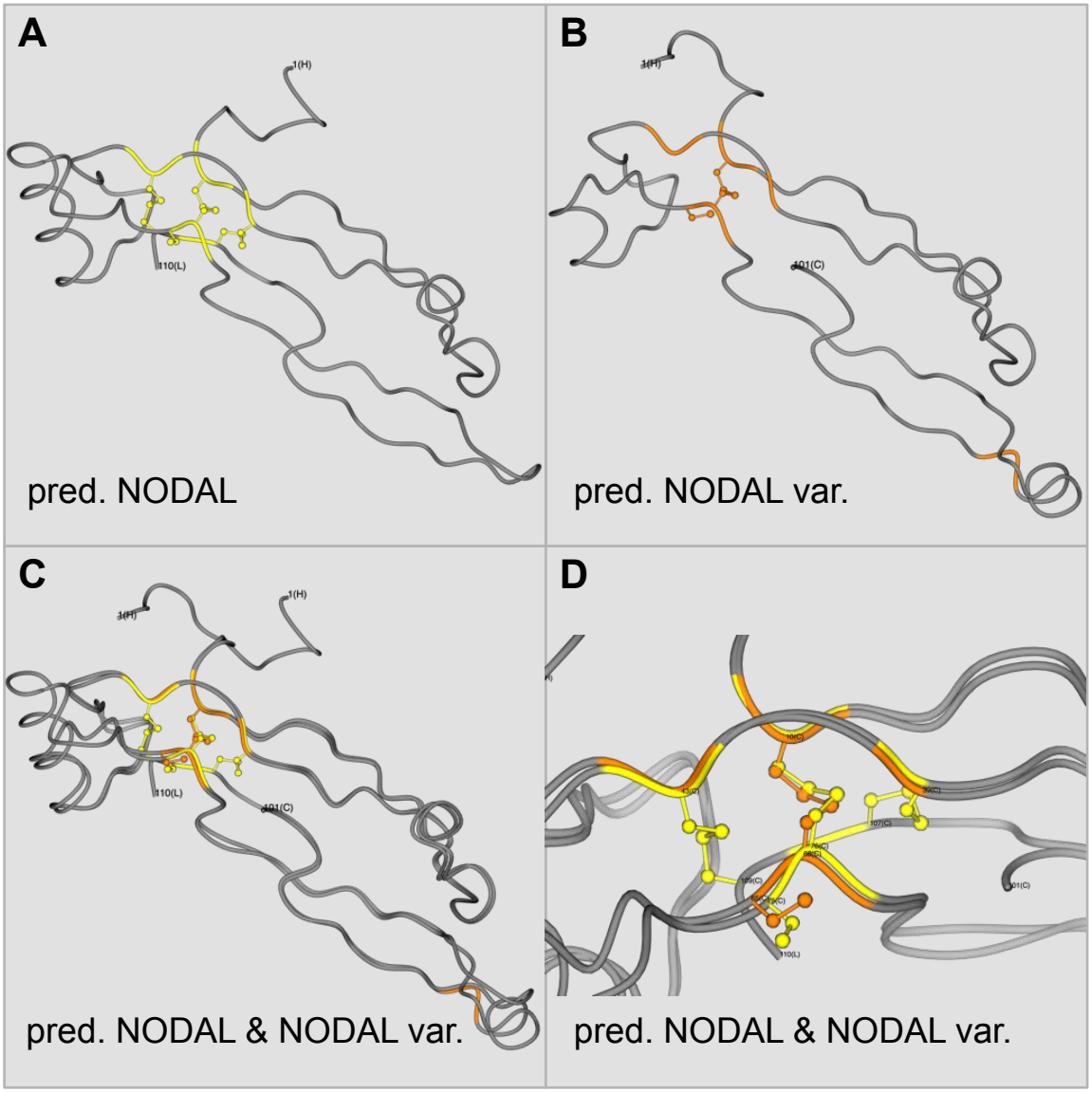


Figure 4.20: Comparison of cysteine arrangements and disulfide bond formation between predicted structures for constitutive NODAL and NODAL variant.

A) and B) Cysteine backbones and side chains are shown in yellow for NODAL and orange for NODAL variant, respectively. The remainder of the peptide backbone is shown in grey. C) Superimposition of predicted structures for NODAL and NODAL variant. D) Magnified view of NODAL cysteine knot region for superimposed structures in C). Cysteines involved in knot structure and putative interchain disulfide bonds are labelled.

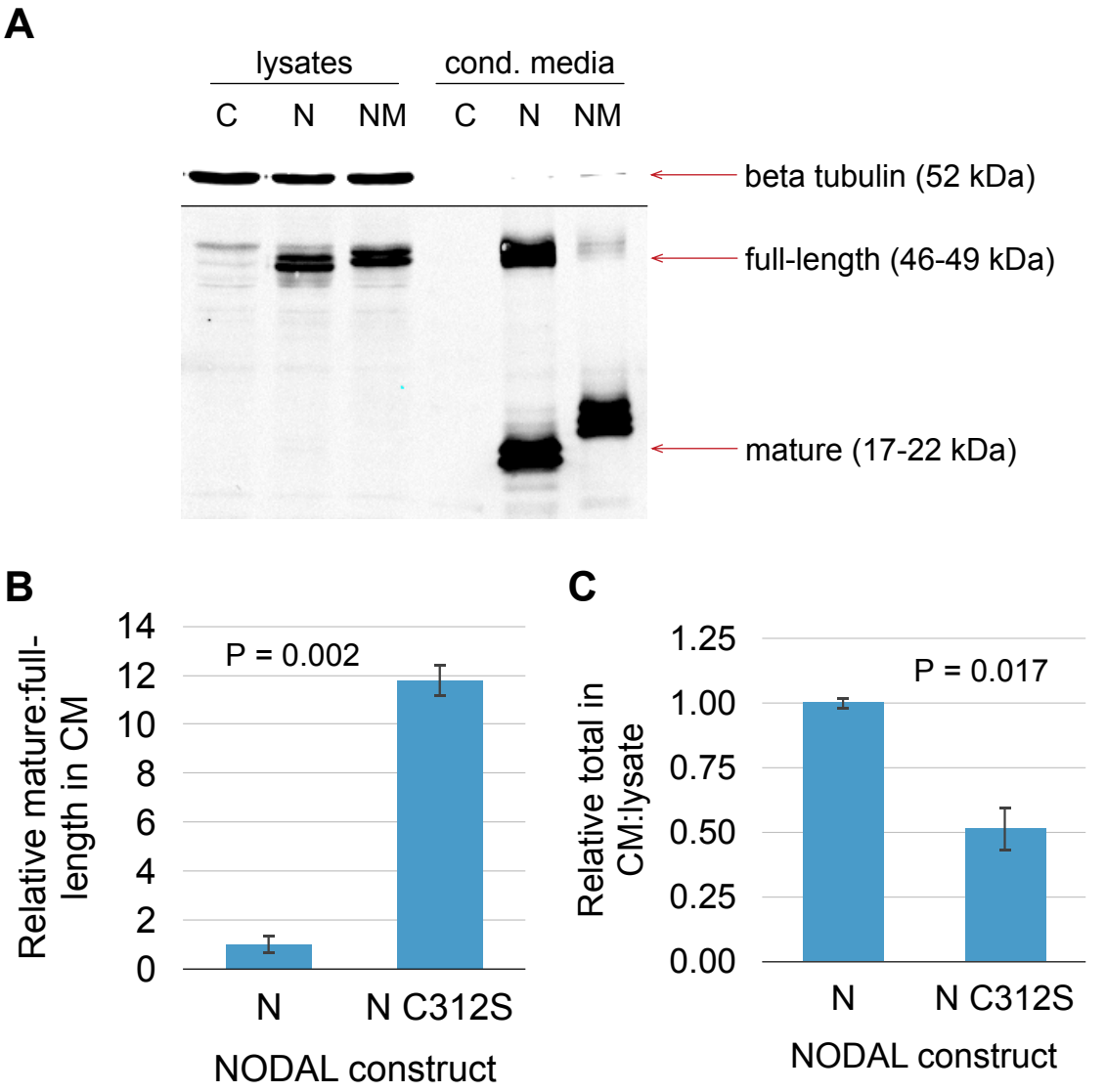


Figure 4.21: Mutation of C312 dramatically affects NODAL processing. Mutation of C312 in the mature domain results in reduced conditioned media:cell lysate ratios of NODAL protein and a dramatic increase in the mature:full-length ratio in conditioned media. Cell lysate from the same number of cells, and conditioned media from the same number of cells were analyzed for each sample. A) "C" = control samples with no NODAL construct. "N" = NODAL. "NM" = NODAL C312S mutant. Approximate sizes of detected bands are shown. B and C) "N" = NODAL, "N C312S" = NODAL C312S mutant. P values are results of significance test for the differences between NODAL constructs using t-tests. Tubulin was included as a loading control. NODAL was detected with an anti-Myc tag antibody. A representative image from two analyses is shown.

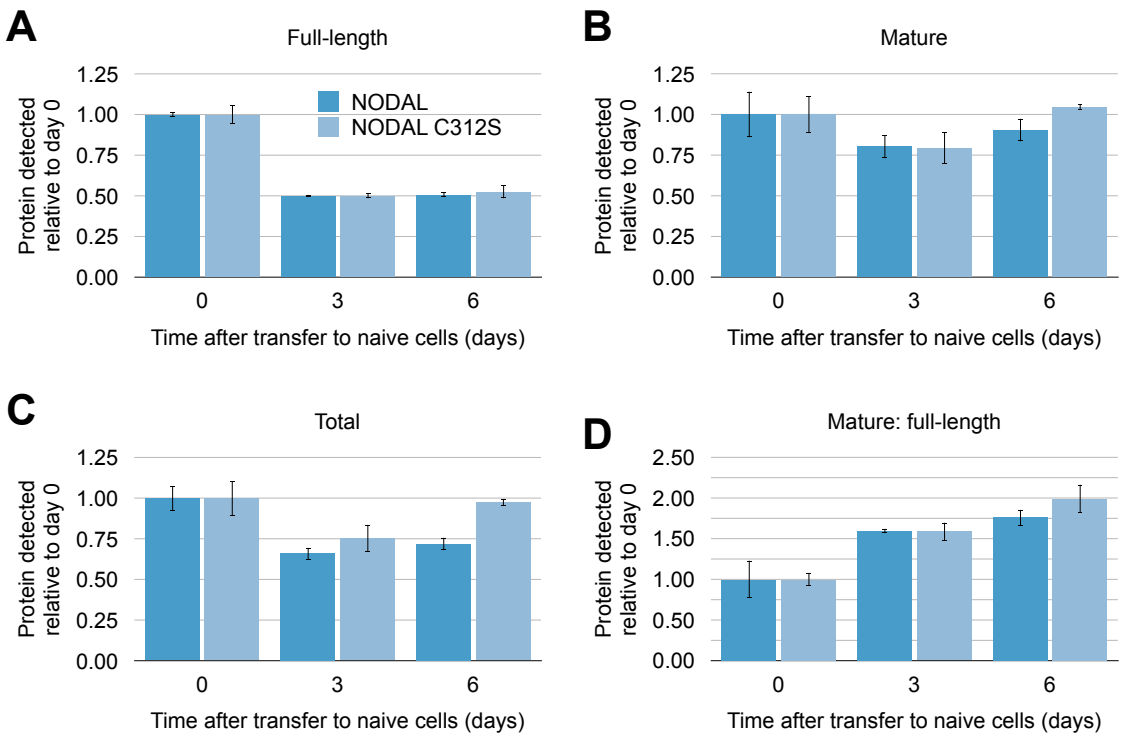


Figure 4.22: Mutation of NODAL C312 had no consistent effect on protein turnover in the media.

Error bars indicate standard deviations. None of full-length, mature, total, or mature:full-length protein showed significant differences across both time points (all $P > 0.05$ by ANCOVA).

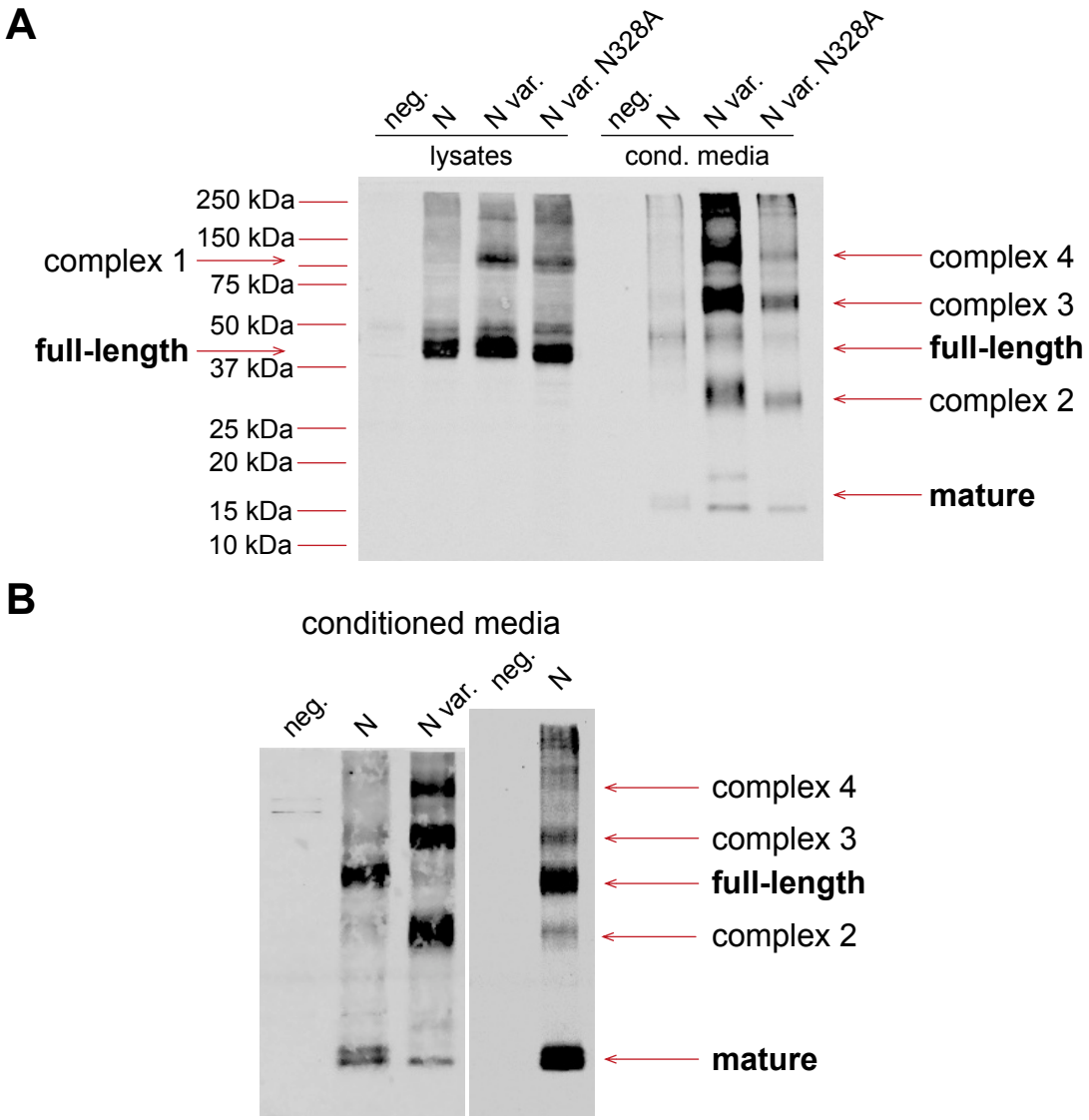


Figure 4.23: Non-reducing analysis of conditioned media reveals less complex formation for NODAL relative to NODAL variant.
"neg." = no NODAL negative sample. "N" = NODAL. "N var" = NODAL variant. "N var. N328A" = NODAL variant N328A mutant. A) Relatively abundant and discrete complexes detected in lysates and conditioned media are indicated. Positions of molecular weight markers are shown. Cell lysate from the same number of cells, and conditioned media from the same number of cells were analyzed for each sample. B) Left: Comparison of NODAL and NODAL variant complexes in conditioned media with equal loading to compensate for increased secretion of NODAL variant (see Figure 4.8). Right: Loading even more NODAL conditioned media reveals low abundance complexes similar to those seen for NODAL variant-conditioned media. Equivalent complexes from A) are indicated.

three additional discrete and abundant complexes, including a possible homodimer complex between 25 and 37 kDa in size (Figure 4.23A). When increased NODAL-conditioned media was loaded to account for differences in abundances between NODAL proteoforms (see Figure 4.8), NODAL complexes similar in size to those formed by the NODAL variant were faintly evident, although much less abundant (Figure 4.23B). Analysis of even higher amounts of NODAL-conditioned media did reveal low abundance complexes more clearly. The low levels of complex formation for NODAL coupled with the even lower abundance of NODAL C312S protein in conditioned media (see Figure 4.21) did not allow for a clear comparison of relative complex formation between the two proteoforms. I next investigated whether the low levels of NODAL complex formation were the result of the tagging strategy used. Structural analysis of the biological assembly formed by the chimera revealed that the C-terminal ends of the mature peptides extend toward the homodimer interface (Figure 4.24). I therefore generated *NODAL* expression constructs that were tagged at the N-terminal end of the mature peptide, which extends away from the homodimer interface, to assess if the C-terminal tagging strategy was confounding analysis of dimerization. Qualitative comparison of *NODAL* constructs tagged at the C-terminus or the N-terminus of the mature peptide (Figure 4.25A) revealed similar patterns of alternative post-translational modification and the presence of processed mature peptides in the conditioned media (Figure 4.25B). However, in comparing MYC versus DYK (Sigma's FLAG) tag, the MYC tag allowed for reliable detection of both full-length and mature NODAL species, while full-length NODAL with an internal DDK tag was undetectable (Figure 4.25B).

Finally, a canonical NODAL signalling assay was used to assess the potential of the NODAL variant to signal relative to constitutive NODAL. Injection of *NODAL* mRNA into single cell stage zebrafish embryos has been shown to induce ectopic *gsc* and *ntl* expression at the shield stage via canonical and Cripto-dependent signalling [22]. Injection of constitutive *NODAL* (n=70) resulted in both gross disruption of gastrulation, and ectopic expression of both *ntl* and *gsc* at the shield stage relative to control (*GFP*)-injected embryos (n=47) (Figure 4.26). Conversely, embryos injected with the *NODAL* variant (n=55) were indistinguishable in their morphological development and expression of *gsc* and *ntl* from both uninjected and control-injected embryos (Figure 4.26).

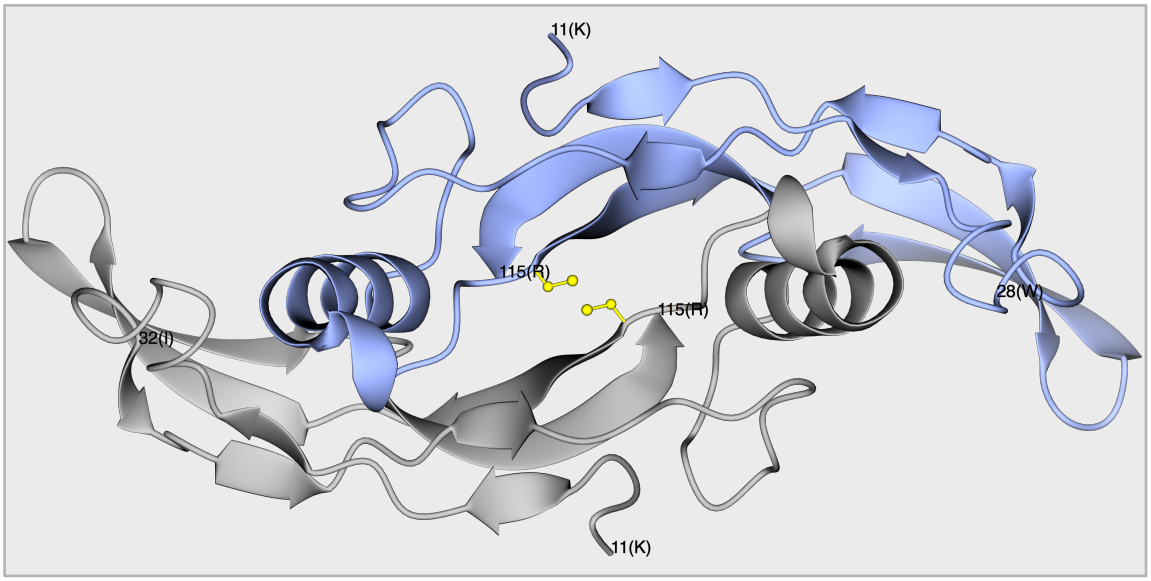


Figure 4.24: Biological assembly of NODAL:BMP2 chimera 4N1D homodimer.

Each monomer mature peptide is coloured differently. N-terminal and C-terminal amino acids are labelled. The side chains of the cysteine analogous to NODAL C312 involved in interchain disulfide bone formation are illustrated in yellow. Note the C-terminus of each subunit (end of beta sheet at 115 R) extending toward the homo-dimerization interface.

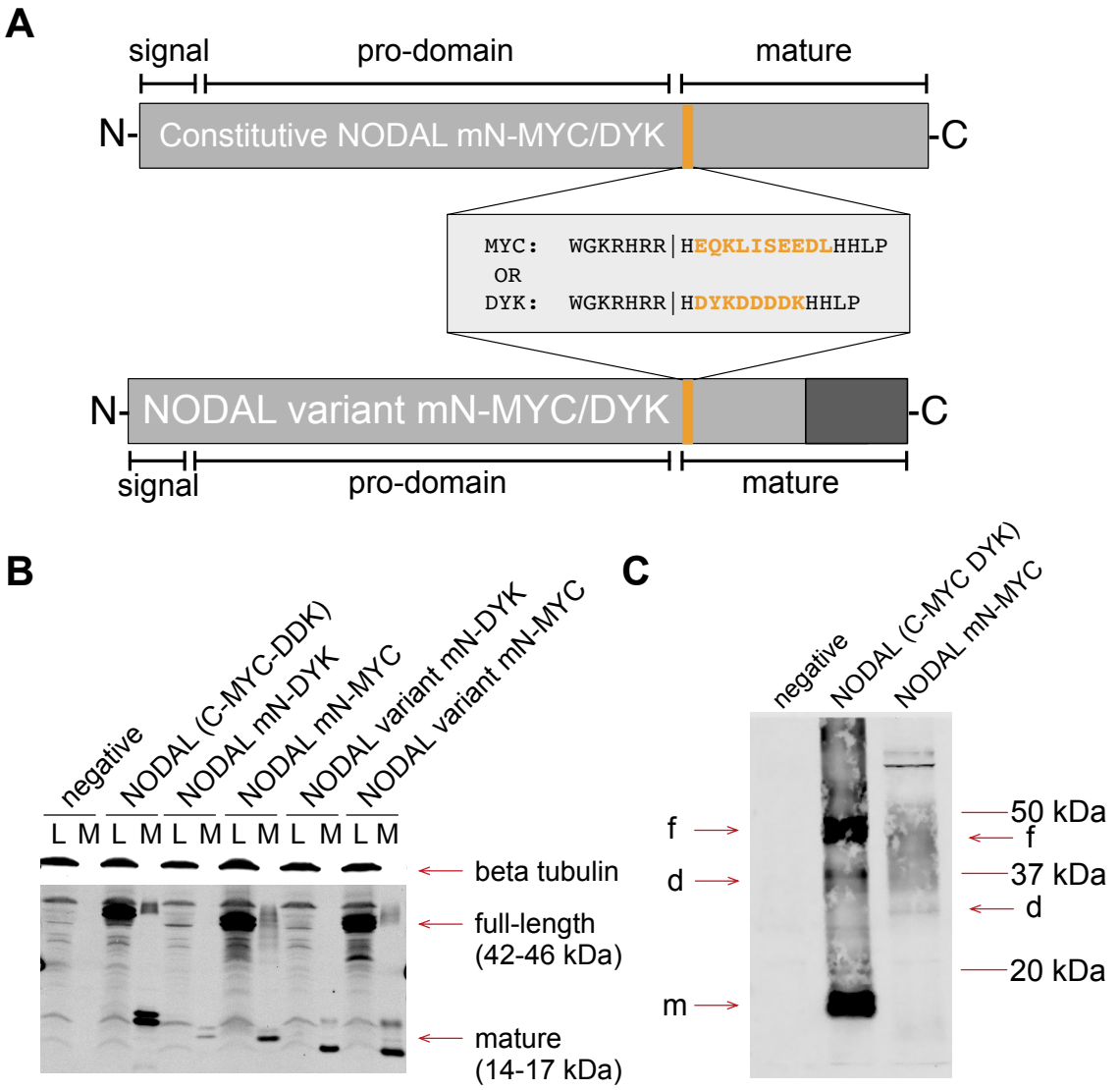


Figure 4.25: Processing and detection of NODAL constructs with different affinity tags.

N-terminal and C-terminal MYC tagging of the NODAL mature peptide produces similar expression profiles in lysates and conditioned media, as well as the presence of mature and full-length peptides. “C-MYC-DYK” = C-terminal dual tag. “mN-MYC” = N-terminal MYC tag of mature NODAL peptide. “mN-DYK” = N-terminal DYK tag of mature NODAL peptide. A) Tag sequence is shown in orange. B) N-terminal DDK tagging of the mature NODAL peptide does not permit efficient detection of full-length protein. Tubulin was included as a loading control. C) Comparative analysis of conditioned media. “f” = full-length protein. “d” = possible NODAL mature homodimer. “m” = mature NODAL peptide. NODAL was detected with an anti-Myc tag antibody.

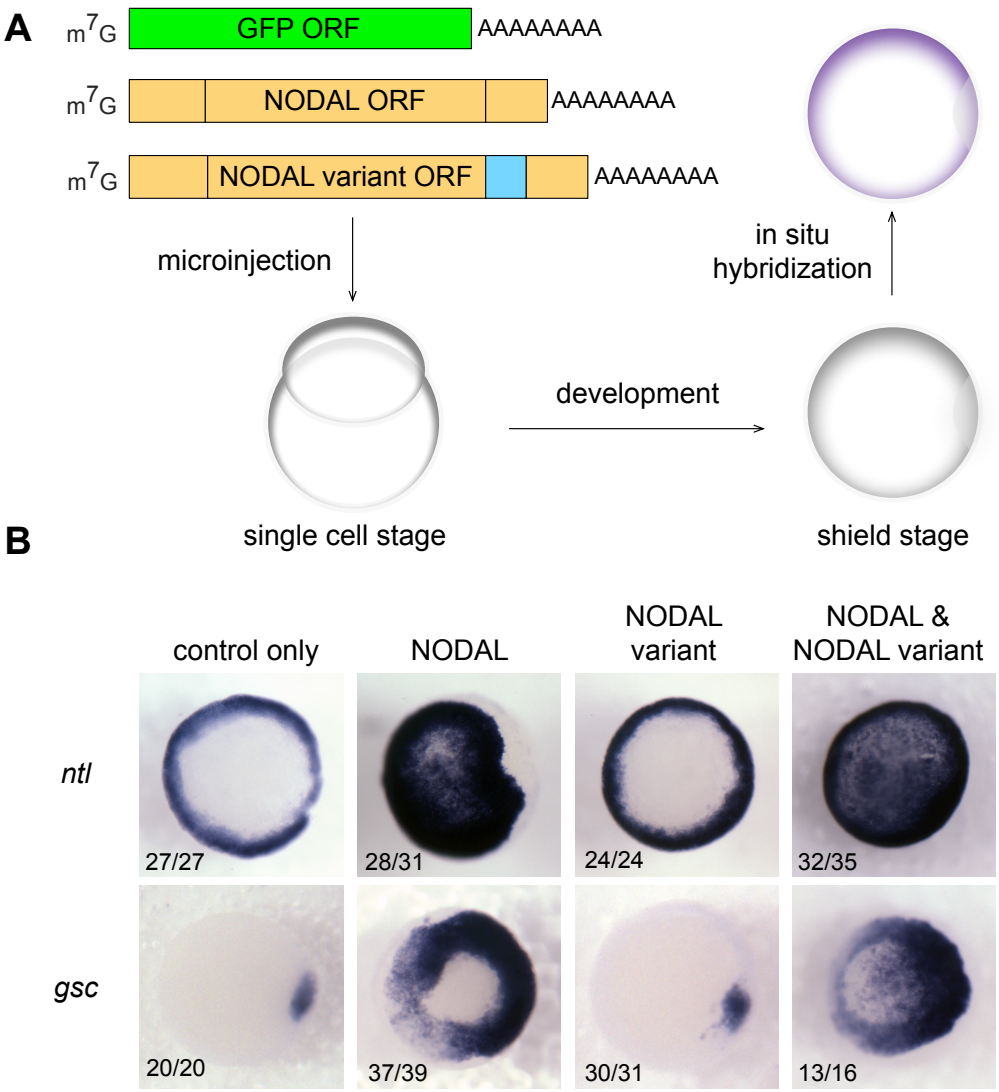


Figure 4.26: Constitutive *NODAL*, but not *NODAL* variant, induces *ntl* and *gsc* expression in a zebrafish model of canonical NODAL signalling. Constitutive NODAL signalling is also not abolished by excess *NODAL* variant. Single cell embryos were injected with mRNA of interest and allowed to develop to shield stage before analysis of *ntl* and *gsc* gene expression. A) "m7G" = 5' 7-Methylguanosine RNA cap. Blue indicates alternative exon; gold indicates constitutive exons. For the single cell stage, a lateral view is shown. For the shield stage, an animal pole view is shown. B) Animal pole views are shown for representative embryos for each condition. Numbers in the bottom left indicate portion of embryos displaying representative expression for each condition. Expression of *ntl* is restricted to the margin in control and *NODAL* variant embryos, and extends to the animal cap in *NODAL* embryos. Expression of *gsc* is restricted to the shield in control and *NODAL* variant embryos, and extends to the margin in *NODAL* embryos.

Specifically, *gsc* expression was restricted to the dorsal organizer (shield) and did not extend around the margin in 20/20 control-injected embryos and 30/31 *NODAL* variant-injected embryos. However, *gsc* expression extended around at least the entire margin and sometimes through the animal cap in 37/39 constitutive *NODAL*-injected embryos. For *ntl*, expression was restricted to the margin for 27/27 control-injected embryos and 24/24 *NODAL* variant-injected embryos, but extended throughout the majority of the animal cap in 28/31 constitutive *NODAL*-injected embryos. I also carried out co-injection of both *NODAL* isoforms to test if the NODAL variant could act as a dominant negative of canonical NODAL signalling. Co-injection did not abolish the NODAL signalling response (Figure 4.26B), suggesting that the NODAL variant is not a potent dominant negative of canonical NODAL signalling in this system.

4.3 Discussion

In presenting the first characterization of a newly identified *NODAL* isoform at the protein level, I have shown that the NODAL variant is biologically distinct from constitutive NODAL. Alternative splicing leads to partial disruption of the TGB-beta superfamily domain and a unique C-terminus that appears to abolish the capacity for canonical NODA*L* signalling by the NODAL variant, likely resulting from disruption of cysteine knot formation that promotes a TGF-beta-like structure. Despite this lack of canonical function, the NODAL variant was efficiently secreted, processed, and stabilized extracellularly in a similar fashion to constitutive NODAL. Moreover, there are definite intriguing differences between the two proteoforms, indicative of consequential biological regulation.

In addition to investigating similarities and divergence between the two NODAL proteoforms, I also modelled molecular aspects of NODAL biology in general. Both NODAL proteoforms were found to be alternatively N-glycosylated. Steady-state expression seemed to favour the proteins with less N-glycosylations, while blocking of de novo translation revealed preferential expression of proteins with N-glycosylations at multiple sites. Proteins with no modified N-glycosylation sites (as in cells treated with tunicamycin) were not detected in untreated cell lysates. This suggests that NODAL is either rapidly N-glycosylated at one site after translation, or that N-glycosylated protein

is preferentially stabilized. My data favours the former model, as completely un-glycosylated NODAL protein was not found to greatly affect turnover of extracellular protein. Intriguingly, while all variant NODAL proteoforms detected in cell lysates contained at least one N-glycosylation, a substantial portion, if not the majority, of secreted and processed NODAL variant was not N-glycosylated, but still displayed relatively high extracellular stability. The data presented here suggest an extremely inefficient secretion of NODAL with no N-glycosylation modifications. The approximate 80% (5-fold) reduction of secreted NODAL relative to that in cell lysates suggests N-glycosylation plays a profound role in the secretion of NODAL. This could be the result of poorly stabilized unmodified NODAL protein due to inefficient folding in the ER or Golgi. The doublet seen for full-length NODAL in cell lysates likely represent differentially N-glycosylated forms of NODAL, although it is possible that an individual N-glycosylation site is differentially processed, as separate mutation of each motif was not performed for constitutive NODAL. It is unclear whether NODAL proteoforms with a second or (in the case of the NODAL variant) third N-glycosylation represent products of a stochastic process, or if their relative abundance is carefully regulated by the cell. That all N-X-S/T sites in the both NODAL proteoforms were likely bona fide sites of N-glycosylation points to active regulation of this process, given the redundancy of N-glycosylation motifs and the fact that the prediction tool used did not identify all putative NODAL sites as strong candidates for true modification.

This work is the first to report differential N-glycosylation of NODAL and directly mutate N-glycosylated amino acid residues. Furthermore, I have characterized an additional novel N-glycosylation site in the mature peptide of the NODAL variant. Putative N-glycosylation sites are found in the mature peptides of various NODAL homologs including several of the Xenopus NODAL-related (Xnr) genes, as well as other mammalian TGF-beta superfamily members such as BMP2,4,6, and 7 [28], and experimental introduction of such N-glycosylation sites enhanced the stability and signalling range of NODAL. However, human constitutive NODAL is not endogenously N-glycosylated in the mature domain. Here we report the existence of such a site endogenously encoded by the novel mRNA sequence in the alternative cassette exon. In contrast to the N-glycosylation previously introduced into the mature NODAL peptide,

this N-glycosylation site did not have a dramatic effect on the processing or stability of extracellular protein, other than a small increase in the mature:full-length peptide ratio over time in the absence of newly translated protein. However, the novel modification did promote increased extracellular protein relative to the constitutively spliced NODAL proteoform, serving as a point of regulation at which the relative abundance of isoforms can differ. Thus, while the *NODAL* variant transcript is expressed at a much lower frequency than the canonical *NODAL* transcript, the relative amount of corresponding secreted protein may triple its relative extracellular abundance. This is an example of how post-transcriptional and post-translational regulation can impact gene expression, and that transcripts with low expression should not automatically be dismissed as functionally unimportant or biological noise.

In addition to alternative N-glycosylation, the processed mature NODAL variant peptide also revealed a doublet for each of the unmodified and N-glycosylated proteoforms. This doublet was similar to that seen for processed constitutive NODAL, and has previously been attributed to o-glycosylation [28]. That the two NODAL proteoforms share N-terminal mature peptide sequence suggests this is likely the region in which NODAL is modified in this fashion.

While N-glycosylation is a typical feature of secreted proteins and dramatically impacted NODAL secretion, I also discovered that mutation of the cysteine involved in interchain disulfide bond formation had a notable impact on both secretion and processing, without negatively impacting extracellular protein stability. Although the C312 residue is in proximity to the cysteine knot, it does not participate in this structure directly and its mutation would not be expected to disrupt the overall structure or folding of the protein. Furthermore, the choice to mutate this cysteine to a serine was based on the desire to introduce a conservative mutation to isolate the effect of interchain disulfide bond formation while minimizing potential structural impact. The findings reported here have the exciting implication that interchain disulfide bond formation, whether homo-dimeric or heterodimeric, is involved in normal secretion and extracellular processing of NODAL. Notably, the mature NODAL peptide with C312S mutation was larger in size

than expected, suggesting the mutation may have introduced a cryptic post-translational modification site which may have potentially confounded this analysis.

Inclusion of paired cell lysates in conditioned media analysis served as an ideal control to account for differences in transfection efficiencies, gene expression levels, and cell number between cells expressing different NODAL proteins. This was of great benefit for analysis of mutant constructs since mutation of key structural residues may affect protein stability in general and thus steady-state expression levels. Thus the differences reported were robust and resulted from the biology of interest, and not technical variability. Relative differences were reported as the absolute levels of protein in the conditioned media and mature:full-length peptide did vary slightly between replicates.

The conditioned media transfer experiments utilized here offered an excellent system to study extracellular NODAL dynamics. The ability to study protein processing in the absence of chemical inhibitors of translation such as cyclohexamide has several advantages. First, the starting quantity of extracellular protein in each cell line can be known precisely. Second, in cyclohexamide-treated cells, additional protein can be secreted into the media after translation proper is inhibited, confounding analysis. Third, cyclohexamide is a global inhibitor of translation, unquestionably resulting in an altered cellular state that may include deregulation of pathways that affect NODAL processing and thus stability. Since NODAL is extracellularly processed, I was not able to study "stability" per se in isolation. Rather, the relative levels of extracellular NODAL are the result of various processing events. In the absence of replenishment by newly translated protein, full-length NODAL may be degraded, internalized, or enzymatically cleaved to yield mature peptide. Similarly, mature peptide may be degraded or internalized.

Other studies have used cleavage resistant mutants and super cleavage mutants to assess the dynamics of full-length and mature NODAL peptides, respectively, in isolation [28]. However, such interventions themselves likely also confound the normal biology and behaviour of the NODAL protein. Furthermore, it is important to consider that the protein forms detected after reducing SDS page do not necessarily represent the actual biological complexes whose dynamics are of interest. For example, for TGFB1, the pro-domain

remains associated with the mature peptide even after proteolytic cleavage [42]. In this study, mature peptides constituting part of an analogous NODAL complex are indistinguishable from “free” mature peptides. Furthermore, in utilizing tagging strategies and antibodies that allow detection of the mature peptide, the dynamics of the cleaved pro-domain in isolation were not investigated. For example, it is possible that mutation of N-glycosylation motifs in the NODAL pro domain has a profound effect on stability of this peptide once cleavage takes place, despite no effect on the dynamics of the full-length precursor protein and the resultant accompanying mature peptide.

In general, the conditioned media transfer experiments reported here suggest that NODAL proteins are quite stable in this system, with over 60% of total protein still detected after six days, and no significant decrease in mature NODAL peptide after three days. These dynamics were much different than those reported by LeGood and colleagues where mature NODAL peptide levels decreased in conditioned media after only ten hours [28]. It is possible that the cells used in the two studies were differentially sensitive to NODAL signals, and that responsiveness to NODAL influences protein turnover. There were also fundamental differences in the two experimental approaches. The study reported here conditioned media for a 48 hour “pulse” allowing high levels of secreted NODAL to accumulate before transfer to untransfected cells for protein breakdown analysis. In contrast, LeGood and colleagues used metabolic labelling to specifically analyze the dynamics of newly secreted NODAL protein for between two and ten hours into fresh media. It is unclear if secretion may be a confounding variable in assessing extracellular stability with this approach. It is possible that higher absolute levels of NODAL or other factors in the conditioned media in our approach had a stabilizing effect. In addition, LeGood and colleagues assessed NODAL uptake and recycling back into the media, which was enhanced with introduction of mature peptide N-glycosylation. Their system utilized pre-incubation of cells on ice before treatment with between 5-fold and 20-fold concentrated conditioned media to facilitate protein uptake. While I did not assess uptake directly in this fashion, no mature NODAL peptides were detected in HEK 293 cell lysates despite high levels of mature peptide in corresponding conditioned media. This was true for all proteoforms analyzed (Figures 4.7, 4.12, and 4.21), and for both C-terminal and N-terminal mature peptide tags (Figure

4.25). These results suggest that spontaneous cellular uptake and stabilization of extracellular NODAL is not very prominent in this experimental system.

It is possible that the tagging strategy used here affected the absolute dynamics of NODAL. However, the same tagging system for each NODAL proteoform investigated ensured that relative differences were robust. C-terminal tagging was chosen to study the dynamics of secretion and processing as mutation of the interface of the pro-domain and mature peptide can impact recognition and cleavage by convertases [43]. However, structural analysis suggests that the N-terminus of the mature NODAL peptide is less structured and extends away from the homodimer interface. Conversely, the C-terminus is part of the cysteine knot in the core of the protein and contributes to the homodimerization interface. Indeed, for constitutive NODAL, a C-terminal tag appeared to reduce complex formation via disulfide bonds relative to an N-terminal tag. However, the NODAL variant still showed high levels of complex formation when a C-terminal tag was used. Regardless, NODAL with a tag at the N-terminus of the mature peptide is likely more suitable for functional study of NODAL, and is utilized for experiments in the next chapter. Notably, full-length NODAL with a mature N-terminal DYK tag was not detected in conditioned media or cell lysates, despite detection of corresponding mature peptide in conditioned media. While this could be the result of “super cleavage", the absence of detection in cell lysates where NODAL is generally not processed suggests a technical problem such as internal epitope masking. Therefore, NODAL with a MYC tag at the N-terminus of the mature peptide was found to have more utility.

The relatively small portion of the conserved TGF-beta superfamily domain disrupted by cassette alternative exon inclusion in the NODAL variant is consistent with a conservative role for alternative splicing in the modulation of conserved domain and whole protein structure [38], as it has been suggested that alternative splicing events leading to substantial domain truncation of large domains are unlikely to result in stable protein products. Recently, an impressive large-scale screen of protein-protein interactions for a collection of human open reading frames revealed functional significance of alternative splicing on a genome wide scale at the protein level [39]. Quantitative analysis of protein-protein interactions for alternatively spliced isoforms

revealed cases with identical, intermediate, and completely distinct interaction profiles. Analysis also revealed that alternatively spliced proteoforms were indistinguishable from protein products of distinct genes in the relatedness of their interaction networks and disease associations.

The observation that the alternatively spliced NODAL proteoforms studied here have different cystine arrangements in their mature peptides and differential capacity for interchain disulfide bond formation raises the possibility that alternative splicing regulates protein-protein interactions for NODAL. In the study by Yang et al, relative to alternatively spliced pairs with similar interaction networks, isoform pairs with the most dramatic "rewiring" of protein-protein interactions were enriched for intrinsically disordered regions, which have been previously identified as frequently modulated by alternative splicing [44-47].

Interestingly, disorder prediction analysis of the two NODAL proteoforms using Pondr-fit [48] revealed similar disorder profiles between the two mature NODAL peptides. Strikingly, unique C-terminal regions of each protein modulated by alternative splicing (residues 103-110 for constitutive NODAL, and residues 92-101 for the NODAL variant, were predicted to be disordered (Appendix B). Since the NODAL C-terminus is known to confer specificity to NODAL signals [22], in addition to interchain disulfide bond formation, it likely plays a key role in receptor-ligand interactions. Another study revealed that alternatively spliced exons with tissue-specific expression were enriched for phosphorylation sites [49]. Although not experimentally assessed here, a scan of Prosite for biologically significant sites and patterns predicted two protein kinase C phosphorylation sites at the C-terminus of the NODAL variant that are absent in constitutive NODAL, suggesting there may be additional modulation of post-translational modification between the two proteoforms beyond the N-glycosylation described here. It is possible that other PTMs such as N-glycosylation are also frequently modulated by alternative splicing in a similar manner on a genome-wide scale. Collectively, these data suggest that alternative splicing is a bona fide mechanism for the modulation of biologically relevant protein function and interaction networks at the protein level. The findings reported here suggest that human *NODAL* may represent a typical case of the

cell's utilization of alternative splicing as a mechanism to modulate protein processing and function, thus expanding the functional proteome.

The zebrafish embryo was used to model canonical NODAL signaling. This model is powerful in that it is a complete and autonomous biological system providing all of the normal developmental context lacking in conventional cell culture models of early development. It also allows for following the spatiotemporal impact of overexpression of a gene of interest, and monitoring for any effect on gross developmental phenotypes. Indeed, injection of constitutive *NODAL* resulted in disruption of gastrulation, as evidenced by irregular rippling of the leading edge of the enveloping layer and a failure of these cells to move toward the vegetal pole during epiboly.

In summary, I have presented evidence that alternatively spliced NODAL proteoforms are distinct in their post-translational modification, secretion from the cell, and their capacity to form protein complexes and signal. These differences are conferred by alternative exon inclusion leading to novel peptide sequence and moderate disruption of the conserved TGF-beta domain and associated cysteine knot motif, likely resulting in a distinct protein structure from that of constitutively spliced NODAL. While the NODAL variant lacked canonical NODAL signalling capacity in a well-regulated non-human embryonic system, the next chapter will explore its functional impact in genetically and epigenetically unstable human cancer models. In these systems, NODAL has been shown to be functionally relevant, however the mechanisms by which it signals are much less well-defined and likely less tightly regulated.

4.4 Methods

4.4.1 Peptide sequence analyses

Pairwise sequence alignments were performed with Emboss needle (http://www.ebi.ac.uk/Tools/psa/emboss_needle/). Extent of conserved domain analysis was conducted using the NCBI Conserved Domain Database CD-search (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Secondary structure predictions were made by JPRED4 (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). N-glycosylation site predictions were performed using the NetNGlyc 1.0 Server

(http://www.cbs.dtu.dk/services/NetNGlyc/). Disulfide bond prediction analysis was conducted using the disulfide bond connectivity prediction option in the DiANNA server (http://clavius.bc.edu/~clotelab/DiANNA/).

4.4.2 NODAL-BMP2 chimera

NODAL-BMP2 chimera NB250 sequence was obtained from the RCSB protein databank (PDB) record for structure 4N1D (http://www.rcsb.org/pdb/explore/explore.do?structureId=4N1D).

4.4.3 Protein structural analysis

All protein structure images and analyses were produced using ccp4/ QtMG molecular graphics software (version 2.10.6; http://www.ccp4.ac.uk/MG/).

4.4.4 Plasmid cloning

A plasmid vector coding for human *NODAL* open reading frame (not including the stop codon) cloned into the pCMV6-Entry vector in frame with a tandem MYC DYK (FLAG) tag (Origene Cat. No. RC211302) was used to over express the constitutive *NODAL* isoform. The equivalent plasmid for the *NODAL* variant was also constructed: The *NODAL* variant open reading frame (not including the stop codon) was cloned from H9 cDNA using the following primers: TATATAGCGATCGCCATGCACGCCCACTGCCTGCC and ATATATACGCGTGCAGACTCTGAGGCTTGGCATGG. The PCR product was digested with AsiSI and MluI (New England Biolabs; Whitby, Ontario, Canada) for insertion into the plasmid backbone. The final construct was sequenced to confirm proper assembly. The pCMV6 plasmid containing a GFP insert was used as a negative control.

Plasmids with internal MYC or DYK tags at the N-terminal end of the mature peptides were constructed as follows: The *NODAL* and *NODAL* variant plasmids were first mutated to eliminate a SalI restriction site upstream of the *NODAL* start codon, and to introduce a stop codon at the C-terminal end of the open reading frame. Synthetic double stranded DNA inserts were synthesized for each tag for each *NODAL* isoform. Inserts were of sufficient length to allow for cloning with SalI and NotI restriction sites flanking

the tag site. Inserts and plasmid backbones were digested with SalI and NotI and subsequently ligated. Final plasmids were sequenced to confirm introduction of desired tags.

4.4.5 Site-directed mutagenesis

Site-directed mutagenesis of *NODAL* plasmids was performed with the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent; Santa Clara, California, USA) according to manufacturer's instructions. Mutated plasmids were sequenced to confirm mutation of desired residues. Mutant codons were chosen to match codons frequently utilized by desired residues in human *NODAL*.

4.4.6 Cell culture and transfection

HEK 293 cells were grown in DMEM supplemented with 10% fetal bovine serum (Gibco/Thermo Fisher; Waltham, Massachusetts, USA) at 37°C with 5% CO_2 supplementation. HEK 293 cells were transfected with desired plasmids using Lipofectamine 3000 (Thermo Fisher) following the manufacturer's protocol. Cells were stably selected with G418 (Thermo Fisher) at 600 μg/mL starting 48 hours after transfection until no parallel mock-transfected cells remained, and then maintained at 100 μg/mL.

4.4.7 Conditioned media

For collection of conditioned media, cells were washed once briefly with PBS, and incubated with excess DMEM at 37°C for one hour before replacement with fresh DMEM to be conditioned. Media was conditioned for 48 hours under standard growth conditions. For each 10 cm culture dish of confluent cells used, 5 mL of media was conditioned. Media was collected and spun at 300 g for 10 minutes to eliminate floating cells and large debris. Remaining media was carefully decanted for concentration or transfer. For concentration, conditioned media was concentrated using Amicon Ultra Centrifugal Filters (Milipore; Billerica, Massachusetts, USA) at 3,000g for approximately 2 hours at 12°C or until media was concentrated in volume by approximately 250-fold. Halt Protease and Phosphatase Inhibitor Cocktail (Thermo Fisher) was added to

concentrated conditioned media. For analysis of NODAL in conditioned media and cell lysates, protein was extracted from cells used to generate conditioned media in parallel.

4.4.8 Protein extraction

Protein was extracted from cells by collecting cells into mammalian protein extraction reagent (mPER; Thermo Fisher) containing the Halt Protease and Phosphatase Inhibitor Cocktail (Thermo Fisher). Lysates were incubated at room temperature for five minutes and mixed thoroughly, then centrifuged at 15,000g for 20 minutes to pellet insoluble cell debris. Protein supernatants were decanted and retained for analysis. Protein concentration was determined using the Pierce BCA Protein Assay Kit (Thermo Fisher) with a standard curve consisting of known concentrations of albumin.

4.4.9 Comparison of cell lysates and conditioned media

For comparison of NODAL levels and processing between cell lysates and conditioned media, samples were isolated from two sets of subsequently generated stable cell lines. An equal number of cells were plated for each stable cell line compared. All samples compared were isolated and analyzed in parallel. For each analysis, cell lysates from an equal number of cells, and conditioned media from an equal number of cells, were analyzed.

4.4.10 Stability experiments

For stability experiments, media was conditioned for 72 hours from one confluent 10 cm dish per stable HEK 293 cell line as described above. On day “minus one” (-1), MDA MB 231 cells were plated in wells of a 12 well plate at approximately 30% confluence. On Day 0, these cells were washed once in PBS, and incubated for one hour in serum-free DMEM at 37°C. Conditioned media from HEK 293 was collected and 1/3 of the media was stored at -80°C to constitute the t=0 sample. The remaining 2/3 of the media was transferred to the recipient cells. On Day 3, 1/2 of the conditioned media on the cells (1/3 of the original conditioned media) was stored at -80°C to constitute the t=1 sample. The remaining conditioned media was also transferred to fresh cells. On Day 6, all of the remaining conditioned media (1/3 of the original conditioned media) was stored at -80°C

to constitute the t=2 sample. Prior to freezing, all samples were spun at 300g for 10 minutes to remove floating cells and large debris. Upon thawing, all samples were concentrated in parallel as described above.

4.4.11 Western blotting

All cell lysate and conditioned media samples were mixed with 4X Laemmli sample buffer (Bio-rad; Hercules, California, USA). For standard reducing analysis, samples were mixed with 5% (v/v) 2-Mercaptoethanol (Sigma-Aldrich; St. Louis, Missouri, USA). For non-reducing analysis, no reducing agent was added. All samples were boiled for five minutes. SDS-PAGE was conducted with 12.5% Acrylamide gels. Precision Plus Protein Dual Color Standards (Bio-rad) were used to confirm approximate molecular weights of detected bands. Proteins were transferred to a low auto fluorescence PVDF membrane (Bio-rad) using the Trans Blot Turbo (Bio-rad) with settings of 25 V and 1.3 A for 15 minutes. After transfer, membranes were washed briefly in PBS, and then blocked for one hour at room temperature with Odyssey Blocking Buffer (Li-Cor; Lincoln, Nebraska, USA). Membranes were incubated overnight in primary antibody solution consisting of Odyssey Blocking Buffer with 0.1% Tween-20 (Sigma-Aldrich). For analysis of NODAL proteoforms with a C-terminal tag, or mature N-terminal MYC tag, mouse anti MYC-tag (9B11) antibody (#2276; Cell Signaling Technologies; Massachusetts, USA) was used at a dilution of 1/1,000. For analysis of mature N-terminal DYK tag proteins, rabbit anti DYK tag antibody (#2368; Cell Signaling Technologies) was used at a dilution of 1/1,000. Rabbit anti β-Tubulin polyclonal antibody (Li-Cor 926-42211) was used at a dilution of 1/1,000 as a loading control for cell lysates. Membranes were then treated with corresponding Li-Cor anti-mouse and anti-rabbit fluorescent secondary antibodies for one hour at room temperature at dilutions of 1/15,000 in Odyssey Blocking Buffer with 0.1% Tween-20 (Sigma-Aldrich) and 0.01% SDS (Thermo Fisher). Membranes were imaged using the Li-Cor Odyssey Clx imaging system. Scans were performed at intensities that did not result in any saturated pixels. Quantification was performed using Li-Cor Odyssey imager software. Notably, this software uses only raw pixel information for quantification, and manipulation of image properties for presentation does not affect quantification.

For stability experiments, four serial dilutions of regularly collected and concentrated conditioned media were included on each gel specific for each stable cell line to constitute a standard curve for protein quantification across the three time points. These samples were prepared at dilutions that were equivalent to 3/2X, 1X, 1/3X, 1/9X, 1/27X, and, where X = input at t=0.

4.4.12 Zebrafish experiments

Constitutive *NODAL* or *NODAL* variant open reading frames were cloned into the pT7TS plasmid (Addgene; Cambridge, Massachusetts, USA; #17091) for *in vitro* transcription using a 5' BglII site and a 3' SpeI site. A control plasmid coding for GFP was also used. These constructs were linearized at a downstream BanHI site and subsequently purified using PureLink PCR Purification Kit (Thermo Fisher). RNA was reverse transcribed from 1 μg of linearized plasmid using the mMessage mMachine T7 *in vitro* transcription kit (Ambion/Thermo Fisher). Transcribed RNA was purified using the RNeasy MinElute Cleanup Kit (Qiagen; Hilden, Germany), and quantified using the Epoch Microplate Spectrophotometer (BioTek; Winooski, Vermont, USA). AB strain zebrafish embryos were injected at the one-cell stage with 250 pg of total RNA diluted in RNase-free water. All injections contained GFP as a positive control, and the total amount of RNA injected was constant for all conditions. Control embryos were injected with only GFP RNA. Embryos were allowed to develop at 28.5°C and monitored until shield stage was reached. Embryos were screened for GFP expression using fluorescence microscopy. Those lacking GFP fluorescence were discarded. Embryos were then fixed in 4% paraformaldehyde for whole mount *in situ* hybridization as previously described [50].

4.5 References

1. Shen, M. M. (2007). Nodal signaling: developmental roles and regulation. *Development*, *134*(6), 1023–1034. doi:10.1242/dev.000166

2. Schier, A. F. (2009). Nodal Morphogens. *Cold Spring Harbor Perspectives in Biology*, *1*(5), a003459–a003459. doi:10.1101/cshperspect.a003459

3. Quail, D. F., Siegers, G. M., Jewer, M., & Postovit, L.-M. (2013). Nodal signalling in embryogenesis and tumourigenesis. *The International Journal of Biochemistry & Cell Biology*, *45*(4), 885–898. doi:10.1016/j.biocel.2012.12.021

4. Bodenstine, T. M., Chandler, G. S., Seftor, R. E. B., Seftor, E. A., & Hendrix, M. J. C. (2016). Plasticity underlies tumor progression: role of Nodal signaling. *Cancer and Metastasis Reviews*, *35*(1), 21–39. doi:10.1007/s10555-016-9605-5

5. Chen, Y.-G. (2009). Endocytic regulation of TGF-beta signaling. *Cell Research*, *19*(1), 58–70. doi:10.1038/cr.2008.315

6. Beck, S., Le Good, J. A., Guzman, M., Ben-Haim, N., Roy, K., Beermann, F., & Constam, D. B. (2002). Extraembryonic proteases regulate Nodal signalling during gastrulation. *Nature Cell Biology*, *4*(12), 981–985. doi:10.1038/ncb890

7. Weiss, A., & Attisano, L. (2012). The TGFbeta Superfamily Signaling Pathway. *Wiley Interdisciplinary Reviews: Developmental Biology*, *2*(1), 47–63. doi:10.1002/wdev.86

8. Tessadori, F., Noël, E. S., Rens, E. G., Magliozzi, R., Evers-van Gogh, I. J. A., Guardavaccaro, D., et al. (2015). Nodal Signaling Range Is Regulated by Proprotein Convertase-Mediated Maturation. *Developmental Cell*, *32*(5), 631–639. doi:10.1016/j.devcel.2014.12.014

9. Guzman-Ayala, M., Ben-Haim, N., Beck, S., & Constam, D. B. (2004). Nodal protein processing and fibroblast growth factor 4 synergize to maintain a trophoblast stem cell microenvironment. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(44), 15656–15660.

10. Eimon, P. M., & Harland, R. M. (2002). Effects of heterodimerization and proteolytic processing on Derrière and Nodal activity: implications for mesoderm induction in Xenopus. *Development*, *129*(13), 3089–3103.

11. Ben-Haim, N., Lu, C., Guzman-Ayala, M., Pescatore, L., Mesnard, D., Bischofberger, M., et al. (2006). The Nodal Precursor Acting via Activin Receptors Induces Mesoderm by Maintaining a Source of Its Convertases and BMP4. *Developmental Cell*, *11*(3), 313–323. doi:10.1016/j.devcel.2006.07.005

12. Osada, S. I., & Wright, C. V. (1999). Xenopus nodal-related signaling is essential for mesendodermal patterning during early embryogenesis. *Development*, *126*(14), 3229–3240.

13. Saloman, D. S., Bianco, C., Ebert, A. D., Khan, N. I., De Santis, M., Normanno, N., et al. (2000). The EGF-CFC family: novel epidermal growth factor-related proteins in development and cancer. *Endocrine Related Cancer*, *7*(4), 199–226.

14. Shen, M. M., & Schier, A. F. (2000). The EGF-CFC gene family in vertebrate development. *Trends in genetics : TIG*, *16*(7), 303–309.

15. Blanchet, M. H., Le Good, J. A., Oorschot, V., Baflast, S., Minchiotti, G., Klumperman, J., & Constam, D. B. (2008). Cripto Localizes Nodal at the Limiting

Membrane of Early Endosomes. *Science Signaling*, *1*(45), ra13–ra13. doi:10.1126/scisignal.1165027

16. Chen, C., & Shen, M. M. (2004). Two Modes by which Lefty Proteins Inhibit Nodal Signaling. *Current Biology*, *14*(7), 618–624. doi:10.1016/j.cub.2004.02.042

17. Liguori, G. L., Borges, A. C., D'Andrea, D., Liguoro, A., Gonçalves, L., Salgueiro, A. M., et al. (2008). Cripto-independent Nodal signaling promotes positioning of the A-P axis in the early mouse embryo. *Developmental Biology*, *315*(2), 280–289. doi:10.1016/j.ydbio.2007.12.027

18. Ding, J., Yang, L., Yan, Y. T., Chen, A., Desai, N., Wynshaw-Boris, A., & Shen, M. M. (1998). Cripto is required for correct orientation of the anterior-posterior axis in the mouse embryo. *Nature*, *395*(6703), 702–707. doi:10.1038/27215

19. Tanaka, C., Sakuma, R., Nakamura, T., Hamada, H., & Saijoh, Y. (2007). Long-range action of Nodal requires interaction with GDF1. *Genes & Development*, *21*(24), 3272–3282. doi:10.1101/gad.1623907

20. Fuerer, C., Nostro, M. C., & Constam, D. B. (2014). Nodal·Gdf1 heterodimers with bound prodomains enable serum-independent nodal signaling and endoderm differentiation. *The Journal of biological chemistry*, *289*(25), 17854–17871. doi:10.1074/jbc.M114.550301

21. Gritsman, K., Zhang, J., Cheng, S., Heckscher, E., Talbot, W. S., & Schier, A. F. (1999). The EGF-CFC Protein One-Eyed Pinhead Is Essential for Nodal Signaling. *Cell*, *97*(1), 121–132. doi:10.1016/S0092-8674(00)80720-5

22. Cheng, S. K., Olale, F., Brivanlou, A. H., & Schier, A. F. (2004). Lefty Blocks a Subset of TGFβ Signals by Antagonizing EGF-CFC Coreceptors. *PLoS Biology*, *2*(2), e30–12. doi:10.1371/journal.pbio.0020030

23. Kelber, J. A., Shani, G., Booker, E. C., Vale, W. W., & Gray, P. C. (2008). Cripto is a noncompetitive activin antagonist that forms analogous signaling complexes with activin and nodal. *Journal of Biological Chemistry*, *283*(8), 4490–4500. doi:10.1074/jbc.M704960200

24. Esquivies, L., Blackler, A., Peran, M., Rodriguez-Esteban, C., Izpisua Belmonte, J. C., Booker, E., et al. (2014). Designer Nodal/BMP2 Chimeras Mimic Nodal Signaling, Promote Chondrogenesis, and Reveal a BMP2-like Structure. *Journal of Biological Chemistry*, *289*(3), 1788–1797. doi:10.1074/jbc.M113.529180

25. Schwarz, F., & Aebi, M. (2011). Mechanisms and principles of N-linked protein glycosylation. *Current Opinion in Structural Biology*, *21*(5), 576–582. doi:10.1016/j.sbi.2011.08.005

26. Blanchet, M.-H., Le Good, J. A., Mesnard, D., Oorschot, V., Baflast, S.,

Minchiotti, G., et al. (2008). Cripto recruits Furin and PACE4 and controls Nodal trafficking during proteolytic maturation. *The EMBO Journal*, *27*(19), 2580–2591. doi:10.1038/emboj.2008.174

27. Mitra, N., Sinha, S., Ramya, T. N. C., & Surolia, A. (2006). N-linked oligosaccharides as outfitters for glycoprotein folding, form and function. *Trends in Biochemical Sciences*, *31*(3), 156–163. doi:10.1016/j.tibs.2006.01.003

28. Le Good, J. A., Joubin, K., Giraldez, A. J., Ben-Haim, N., Beck, S., Chen, Y., et al. (2005). Nodal Stability Determines Signaling Range. *Current Biology*, *15*(1), 31–36. doi:10.1016/j.cub.2004.12.062

29. Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., et al. (2012). The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science*, *338*(6114), 1587–1593. doi:10.1126/science.1230612

30. Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, *40*(12), 1413–1415. doi:10.1038/ng.259

31. Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. doi:10.1038/nature07509

32. Hamid, F. M., & Makeyev, E. V. (2014). Emerging functions of alternative splicing coupled with nonsense-mediated decay. *Biochemical Society Transactions*, *42*(4), 1168–1173. doi:10.1042/BST20140066

33. Lareau, L. F., & Brenner, S. E. (2015). Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. *Molecular biology and evolution*, *32*(4), 1072–1079. doi:10.1093/molbev/msv002

34. Lareau, L. F., Inada, M., Green, R. E., Wengrod, J. C., & Brenner, S. E. (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, *446*(7138), 926–929. doi:10.1038/nature05676

35. Ni, J. Z., Grate, L., Donohue, J. P., Preston, C., Nobida, N., O'Brien, G., et al. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes & Development*, *21*(6), 708–718. doi:10.1101/gad.1525507

36. de Klerk, E., & t Hoen, P. A. C. (2015). Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends in Genetics*, *31*(3), 128–139. doi:10.1016/j.tig.2015.01.001

37. Smith, L. M., Kelleher, N. L., Linial, M., Goodlett, D., Langridge-Smith, P., Ah Goo, Y., et al. (2013). Proteoform: a single term describing protein complexity. *Nature Methods*, *10*(3), 186–187. doi:10.1038/nmeth.2369

38. Hegyi, H., Kalmar, L., Horvath, T., & Tompa, P. (2011). Verification of alternative splicing variants based on domain integrity, truncation length and intrinsic protein disorder. *Nucleic Acids Research*, *39*(4), 1208–1219. doi:10.1093/nar/gkq843

39. Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., et al. (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*, *164*(4), 805–817. doi:10.1016/j.cell.2016.01.029

40. Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., & Stamm, S. (2013). Function of alternative splicing. *Gene*, *514*(1), 1–30. doi:10.1016/j.gene.2012.07.083

41. Galat, A. (2011). Common structural traits for cystine knot domain of the TGFβ superfamily of proteins and three-fingered ectodomain of their cellular receptors. *Cellular and Molecular Life Sciences*, *68*(20), 3437–3451. doi:10.1007/s00018-011-0643-4

42. Dijke, ten, P., & Arthur, H. M. (2007). Extracellular control of TGFβ signalling in vascular development and disease. *Nature reviews. Molecular cell biology*, *8*(11), 857–869. doi:10.1038/nrm2262

43. Constam, D. B., & Robertson, E. J. (1999). Regulation of bone morphogenetic protein activity by pro domains and proprotein convertases. *The Journal of Cell Biology*, *144*(1), 139–149.

44. Buljan, M., Chalancon, G., Eustermann, S., Wagner, G. P., Fuxreiter, M., Bateman, A., & Babu, M. M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Molecular Cell*, *46*(6), 871–883. doi:10.1016/j.molcel.2012.05.039

45. Ellis, J. D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J. A., et al. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular Cell*, *46*(6), 884–892. doi:10.1016/j.molcel.2012.05.037

46. Haynes, C., Oldfield, C. J., Ji, F., Klitgord, N., Cusick, M. E., Radivojac, P., et al. (2006). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS computational biology*, *2*(8), e100. doi:10.1371/journal.pcbi.0020100

47. Weatheritt, R. J., Davey, N. E., & Gibson, T. J. (2012). Linear motifs confer functional diversity onto splice variants. *Nucleic Acids Research*, *40*(15), 7123–

7131. doi:10.1093/nar/gks442

48. Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K., & Uversky, V. N. (2010). PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1804*(4), 996–1010. doi:10.1016/j.bbapap.2010.01.011

49. Merkin, J., Russell, C., Chen, P., & Burge, C. B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, *338*(6114), 1593–1599. doi:10.1126/science.1228186

50. Drummond, D. L., Cheng, C. S., Selland, L. G., Hocking, J. C., Prichard, L. B., & Waskiewicz, A. J. (2013). The role of Zic transcription factors in regulating hindbrain retinoic acid signaling. *BMC developmental biology*, *13*(1), 31. doi:10.1186/1471-213X-13-31

Chapter 5

5 Differential effects of *NODAL* isoforms on cancer phenotypes, and improving *NODAL* modelling using precision genome editing.

5.1 Introduction

Ovarian cancer is the 5th most commonly diagnosed cancer for women and the deadliest cancer affecting the female reproductive system. Approximately 70-80% of all cases are classified as high-grade serous epithelial ovarian carcinoma [1]. While *NODAL* is known to be aberrantly expressed in numerous cancers (reviewed in [2, 3]), only a limited amount of work has investigated *NODAL* expression and function in ovarian carcinoma [4, 5]. In patient samples of several different cancers, *NODAL* expression is positively correlated with disease stage. Consistent with these clinical observations, NODAL has been shown to promote and maintain pro-tumourigenic phenotypes in numerous in vitro models of these cancers. Such phenotypes include increased cellular proliferation and increased xenograft tumour volume in mice, as well as heightened stem cell-like properties such as anchorage independent growth and spheroid formation (reviewed in [2, 3]). However, in limited studies of ovarian cancer cell lines, *NODAL* over-expression was shown to promote decreased proliferation and increased apoptosis [5]. Interestingly, using a PCR assay that crossed the constitutive exon 1- constitutive exon 2 junction, detectable levels of *NODAL* were reported for several ovarian cancer cell lines including A2780S (sensitive) and A2780CP (resistant) cells commonly used to model ovarian cancer resistance to standard platinum-based chemotherapies [5]. *NODAL* expression has also been shown to increase in response to cisplatin treatment [6]. Perhaps surprisingly, a role for NODAL in conferring resistance to chemotherapy has not yet been reported, although inhibition of NODAL has sensitized cells to chemotherapy in models of pancreatic cancer and melanoma [7, 8].

The fact that NODAL has been found to play a pro-tumourigenic role in most cancers, but has so far been shown to promote anti-tumourigenic phenotypes in ovarian cancer models is just one example of how cancer heterogeneity poses a challenge to

experimental modelling of human cancers. Beyond inter-cancer and inter-tumour heterogeneity, heterogeneity within a tumour or cell line is also a confounding factor. Subcultures of a cell line can drift over time and diverge phenotypically due to intrinsic genetic and epigenetic instability [9-11]. It is possible that this effect is even more profound if antibiotic selection is applied for enrichment of efficiently transfected cells. Pertinent to modelling NODAL activity, there are also potential dose or duration-dependent effects of NODAL signals [12, 13]. Moreover, specific aspects of *NODAL* biology cannot always be gleaned from more strictly regulated developmental systems. Even though cancers frequently hijack developmental programs, the engaged signal transduction pathways are not always subject to the endogenous regulation they would experience in normal biological systems. For example, in embryonic systems, expression of endogenous NODAL inhibitors such as the Lefty proteins [14] is induced upon activation of NODAL signalling, providing negative feedback that helps restrict NODAL activity and signalling range [15-17]. However, low or undetectable levels of Lefty have been reported in several cancers in both patient samples and cell lines [18, 19], suggesting that NODAL is not always subject to this mechanism of endogenous inhibition in cancer. Collectively, these factors present challenges to the study of *NODAL* in human cancer. As such, the development of robust experimental models to study *NODAL* in cancer will be paramount to investigating potential differences in *NODAL* function between cancers.

Precision genome editing offers numerous opportunities for robust modelling of gene function through the introduction stable mutations at genomic targets of interest. To date, genome editing has not been used to study *NODAL* gene function. The advent of the transcription activator-like effector nuclease (TALEN) and more recently, clustered regularly interspaced short palindromic repeats (CRISPR-Cas) systems have accelerated the adoption of precision genome editing in many fields of molecular biology (reviewed in [20-23]). TALEN nucleases consist of a modular DNA binding domain fused to an endonuclease domain to induce double-stranded breaks at a DNA target. CRISPR-Cas systems rely on base pairing between an exogenous guide RNA (gRNA) and an endogenous genomic DNA (gDNA) target to deliver the CRISPR-associated (Cas) endonuclease. Regardless of the genome editing system used, the generation of a double

stranded break (DSB) is repaired by the cell using either the non-homologous end joining (NHEJ) pathway, or homology-directed repair (HDR). NHEJ is error-prone and frequently generates small indel mutations [24]. Such mutations are easily exploited for functional gene knock-out studies. Induction of a DSB at a desired target can also dramatically improve donor integration for targeted stable integration of exogenous DNA.

Genome editing is already transforming how molecular biology research is conducted. The ability to precisely modify the genome in a targeted fashion has numerous applications relevant to cancer. These include the functional knockout of endogenous genes or alleles such as tumour suppressors or oncogenes, and the targeted introduction or correction of specific acquired mutations or inherited polymorphisms. In addition to robust modelling of genetic contributions to cancer cell function *in vitro*, genome editing has obvious potential as a therapeutic tool in the treatment of cancers [25]. Indeed, the first approved use of CRISPR technology in a clinical trial involves *ex vivo* editing of T cells in an effort to enhance their efficacy and longevity in cancer immunotherapy [26].

For the promise of precision genome editing technologies to be realized, it is also important that reliable screening methods are available for detection of desired mutations. Such assays need to be quantitative, specific, sensitive, and universal in that they can be readily adapted to any target of interest. Genome editing experiments often result in low mutation frequencies in bulk populations of treated cells. Therefore, precise quantification of mutation rates is extremely important for optimization of genome editing protocols and downstream workflow, such as determining how many single cell-derived clones to screen for desired mutations.

While next generation sequencing offers a gold standard for quantitative determination of nuclease-induced mutation detection, such approaches are often not practical. Several different methods to screen for nuclease-induced mutations have been reported [27-30]. However, the most widely used assays to screen for mutations utilize the so-called “mismatch nucleases” T7E1 or “Surveyor” that recognize and cleave heteroduplexed DNA amplicons containing mismatched base-pairs [31]. These assays have several

shortcomings: First, they require a relatively large amount of starting material to generate sufficient levels of purified PCR product corresponding to the target locus. This requirement does not allow for rapid workflows, as significant cellular expansion is needed after enriching for edited cells using selection or sorting, and again after the generation of single cell-derived clones. Second, there are obvious limitations for sensitivity, as digested fragments that do not make up a large portion of the total amplified target molecules are hard to distinguish from background noise on an electrophoretic gel. Furthermore, targets that cannot be efficiently amplified may not result in bright bands. Indeed, due to the nature of intercalating DNA stains, each digested fragment loses a minimum of 50% of its signal relative to its parent band. Third, this method has very limited utility for screening of single-cell derived clones. For a typical diploid target locus, a clone with both alleles successfully mutated by NHEJ, but containing distinct indels, will be indistinguishable from a clone with one mutated allele and one wild type allele, as each of these samples would contain a 50-50 mix of distinct alleles. In both cases, approximately half of the duplexed DNA would be in the heteroduplexed form. Fourth, these assays generally require the generation of amplicons of at least 400 base pairs to ensure digested fragments are of sufficient length to be visualized. This increases the chances of the amplicon encompassing a polymorphism that is heterozygous in the sample or cell line being used. An endogenous heterozygous SNP or mutant allele anywhere in the amplicon can be recognized by the nuclease and lead to a false-positive signal, even in unedited cells. This is especially problematic in cancer cell lines and samples where mutation frequencies are extremely high and are often unknown for a particular locus of interest.

The emergence of droplet digital PCR (ddPCR) technologies provide a new opportunity for mutation screening that provides superior sensitivity, is absolutely quantitative, and can easily be adapted to any target of interest. Detection of NHEJ-induced indels as well as donor-derived mutations of interest using ddPCR has just recently been reported [32, 33]. Due to the ability to obtain absolute quantifications from very small amounts of DNA, this methodology holds great promise as a preferred method of screening. However, the utility and performance of such assays have not yet been thoroughly assessed.

In this chapter, we examine the impact of over expression of both *NODAL* isoforms on cellular response to carboplatin chemotherapy in A2780S ovarian cancer cells. This is used as a case study to highlight potential drawbacks of over-expression studies, and a motivation to develop robust models for *NODAL* function in cancer cells using precision genome editing. In addition to generating two such models, we also develop tools to streamline cloning and ddPCR mutation screening assays that together improve genome editing workflows.

5.2 Results

To compare the functional impact of the two human *NODAL* splice variants in human cancer, we conducted over-expression studies in A2780S ovarian carcinoma cells. Since A2780S cells have been used as a model to study ovarian carcinoma resistance to chemotherapy, we were interested in testing if *NODAL* expression could confer resistance to the carboplatin, a drug commonly included in standard chemotherapy regimens for patients with epithelial ovarian carcinoma. Relative to stably selected control cells expressing GFP, over expression of constitutive *NODAL* resulted in a 33% increase in total cell metabolism according to an MTT assay when cells were treated with between 3.1 and 12.5 μM of carboplatin ($P < 0.01$; Figure 5.1A). Over-expression of the *NODAL* variant resulted in only a 6% increase in total cell metabolism, and this increase was not statistically significant relative to control cells (Figure 5.1A). We next tested the impact of stable *NODAL* isoform expression on colony formation potential in the absence of any chemotherapy. Interestingly, both constitutive *NODAL* and the *NODAL* variant promoted increased colony forming capacity relative to control cells. (Figure 5.1B). This result prompted us to test clonogenic viability of A2780S cells after treatment with carboplatin. In contrast to the response of all cells to carboplatin treatment, clonogenic growth potential was an average of 28-fold higher for *NODAL*-expressing cells and an average of 6-fold higher for *NODAL* variant-expressing cells when treated with between 3.1 and 12.5 μM carboplatin (Figure 5.1C).

Based on these findings, we were interested in determining if *NODAL*-expressing cells displayed altered expression of genes known to be involved in cancer cell drug resistance. Using a SYBR-green real time PCR array, *NODAL* expressing cells were found to

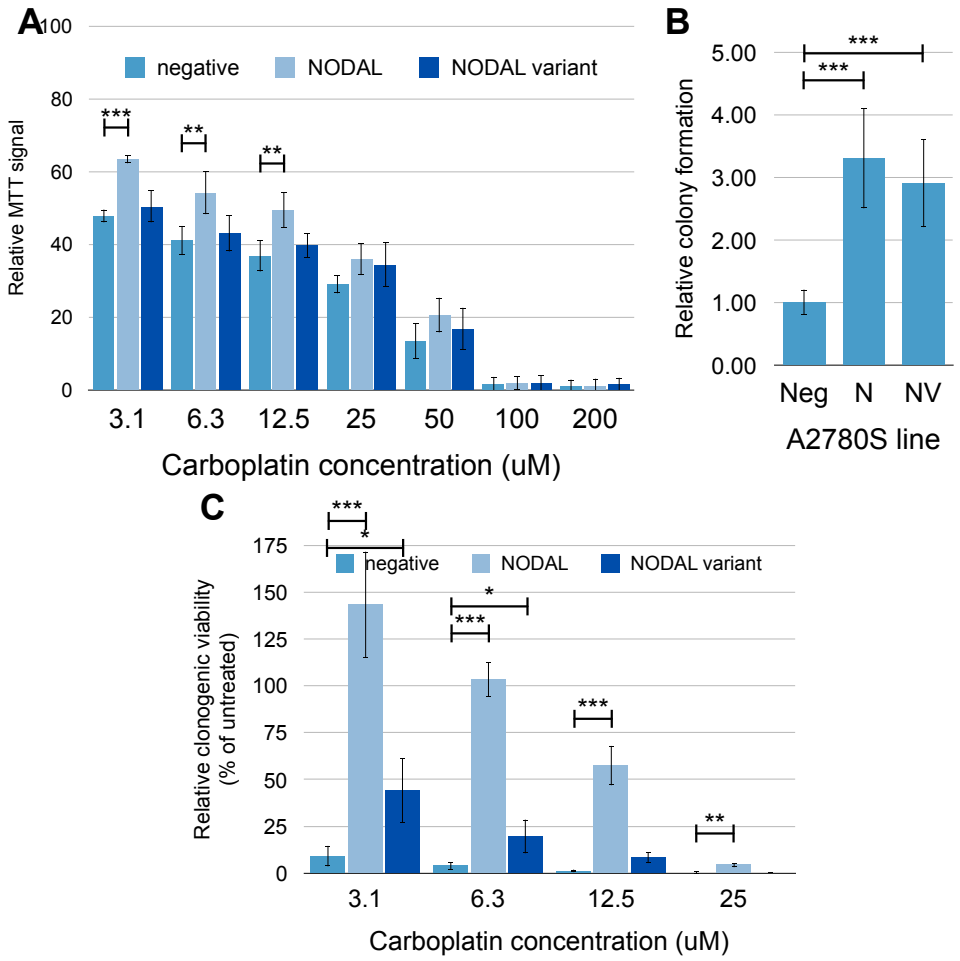


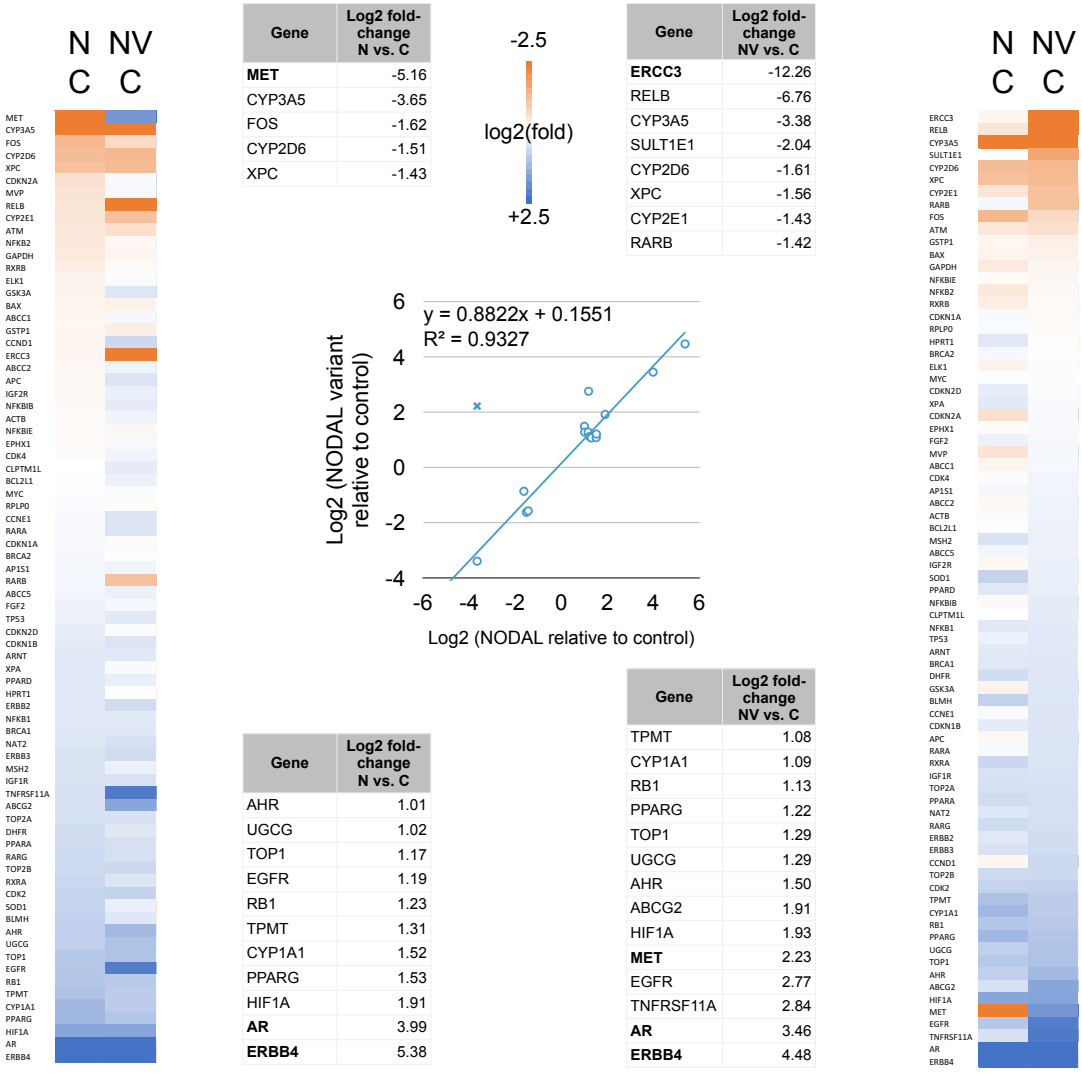
Figure 5.1: Over-expression of *NODAL* and *NODAL* variant isoforms in A2780S ovarian carcinoma cells.

NODAL isoform overexpression resulted in increased colony formation capacity and conferred differential resistance to carboplatin chemotherapy. A) Constitutive *NODAL* over expression results in slightly increased total cell metabolism after treatment with carboplatin relative to control cells, while over-expression of *NODAL* variant does not. B) Both *NODAL* and *NODAL* variant over-expression results in increased colony formation capacity of A2780S cells in the absence of drug treatment. “Neg” = GFP-expressing negative control cells. “N” = *NODAL* over-expression. “NV” = *NODAL* variant over-expression. C) Both *NODAL* and *NODAL* variant confer increased capacity for clonogenic growth in the presence of carboplatin, with both the magnitude and range of this effect extended for constitutive *NODAL* relative to the *NODAL* variant. Error bars indicate standard deviations for three experiments. Asterisks indicate results of ANOVA tests of statistical significance, with “*” = $P < 0.05$, “**” = $P < 0.01$, and “***” = $P < 0.001$.

display altered (> 2-fold change) gene expression for 16 of 73 genes (22%) tested relative to control cells (Figure 5.2). Similarly, *NODAL* variant expressing cells displayed differential expression for 21 of 74 genes (28%) tested (Figure 5.2). Except for one outlying gene (*MET*), the changes in gene expression profiles for both *NODAL* and *NODAL* variant were extremely similar (coefficient of determination (R^2) = 0.9327 when *MET* outlier was excluded). Of the genes with differential expression between *NODAL* expressing and control cells, 15 of the 17 were also differentially expressed (>2-fold) in the same direction in *NODAL* variant expressing cells (Figure 5.2).

We next sought to validate changes in expression for several target genes from the PCR array with independent primer probe assays in droplet digital PCR (ddPCR). *AR* (Androgen receptor) and *ERBB4* (V-erb-a erythroblastic leukemia viral oncogene homolog 4) were selected as they were the two most highly upregulated genes by both *NODAL* and the *NODAL* variant. *MET* (Met proto-oncogene/ hepatocyte growth factor receptor) was selected as it was the most down-regulated gene upon *NODAL* expression, but unchanged in *NODAL* variant-expressing cells, *ERCC3* (Excision repair cross-complementing rodent repair deficiency, complementation group 3) was selected as it was the most differentially expressed gene between *NODAL* variant and *NODAL* expressing cells, and *EGFR* (Epidermal growth factor receptor) was selected as it was differentially upregulated by the *NODAL* variant relative to *NODAL* (Figure 5.2). Although the exact magnitudes differed, results of ddPCR assays confirmed *NODAL*-induced changes in expression for each of *AR*, *ERBB4*, *MET*, and *EGFR*. Changes in *ERCC3* expression were not confirmed, and it was not analyzed further (Figure 5.3).

Next, a second independent set of stable cell lines was generated to assess to what degree the *NODAL* isoforms consistently induced robust changes in expression of these genes. These are referred to as “set 2,” and the original set of stable cell lines as “set 1.” Both *EGFR* and *MET* were more highly expressed in *NODAL* variant than *NODAL* expressing cells for both replicates (Figure 5.4A). *AR* was again induced by *NODAL*, but *ERBB4* was not (Figure 5.4B). The *NODAL* variant again induced *ERBB4*, but not *AR*. In fact, in the newly generated stable cells, *AR* expression was dramatically reduced in *NODAL* variant expressing cells relative to control cells (Figure 5.4C).



Gene	Log2 fold-change N vs. C
MET	-5.16
CYP3A5	-3.65
FOS	-1.62
CYP2D6	-1.51
XPC	-1.43

Gene	Log2 fold-change NV vs. C
ERCC3	-12.26
RELB	-6.76
CYP3A5	-3.38
SULT1E1	-2.04
CYP2D6	-1.61
XPC	-1.56
CYP2E1	-1.43
RARB	-1.42

Gene	Log2 fold-change N vs. C
AHR	1.01
UGCG	1.02
TOP1	1.17
EGFR	1.19
RB1	1.23
TPMT	1.31
CYP1A1	1.52
PPARG	1.53
HIF1A	1.91
AR	3.99
ERBB4	5.38

Gene	Log2 fold-change NV vs. C
TPMT	1.08
CYP1A1	1.09
RB1	1.13
PPARG	1.22
TOP1	1.29
UGCG	1.29
AHR	1.50
ABCG2	1.91
HIF1A	1.93
MET	2.23
EGFR	2.77
TNFRSF11A	2.84
AR	3.46
ERBB4	4.48

Figure 5.2: *NODAL* and *NODAL* variant over-expression induce similar gene expression profiles for genes related to drug resistance in cancer cells. Heat maps are shown for differential gene expression between either *NODAL* ("N"; left) and *NODAL* variant ("NV"; right) over-expressing cells relative to control cells. Genes are sorted from most decreased to most increased (top to bottom), by *NODAL* for the left heat map, and *NODAL* variant for the right heat map. Genes with fold changes > 2 (log2(fold-change > 1) are shown in table form. Bold gene symbols indicate genes selected for ddPCR validation and follow-up.

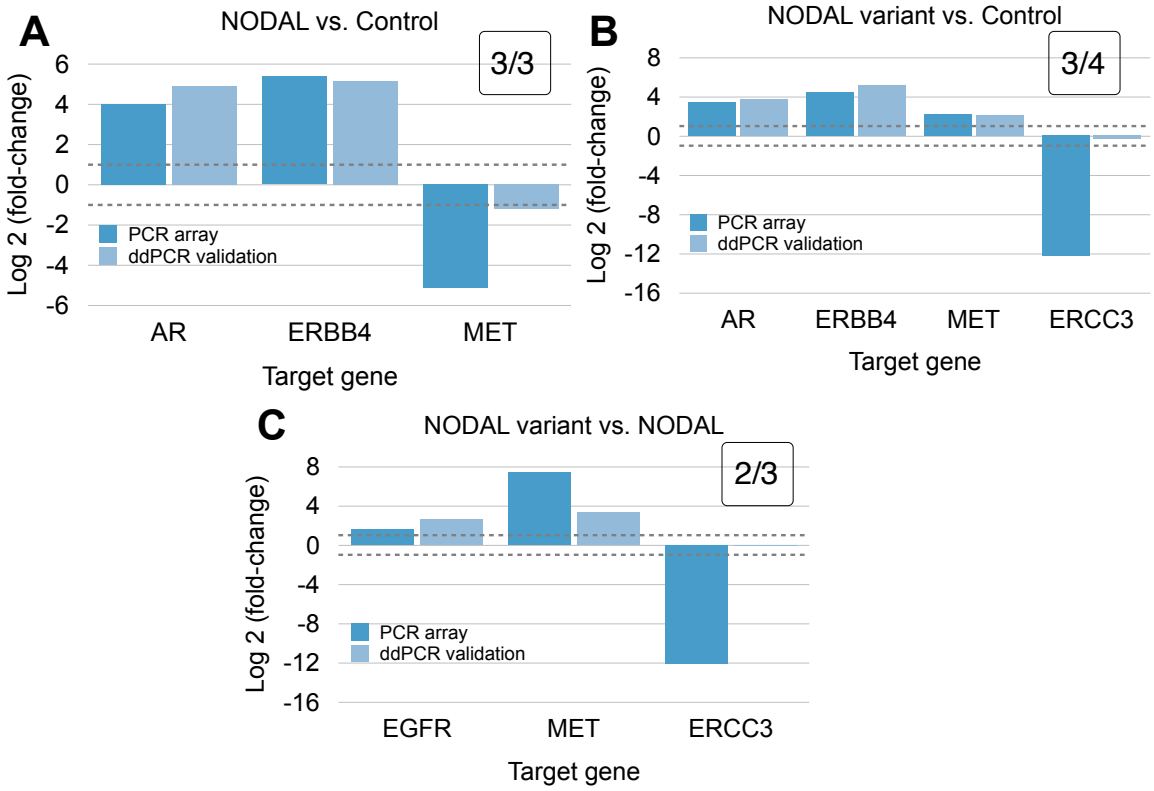


Figure 5.3: Validation of select differentially expressed genes from PCR array with independent ddPCR assays.
Differences in gene expression for each pair of stable cell lines is shown. Changes in expression of all genes selected for follow up except for ERCC3 were validated. Box in the upper right hand corner of each chart indicates the number of genes with differential gene expression (fold-change > 2) according to both SYBR green real time PCR array and ddPCR assays. Dashed grey lines indicate threshold for differential gene expression (fold change of > 2; log2(fold-change) > 1) in either direction.

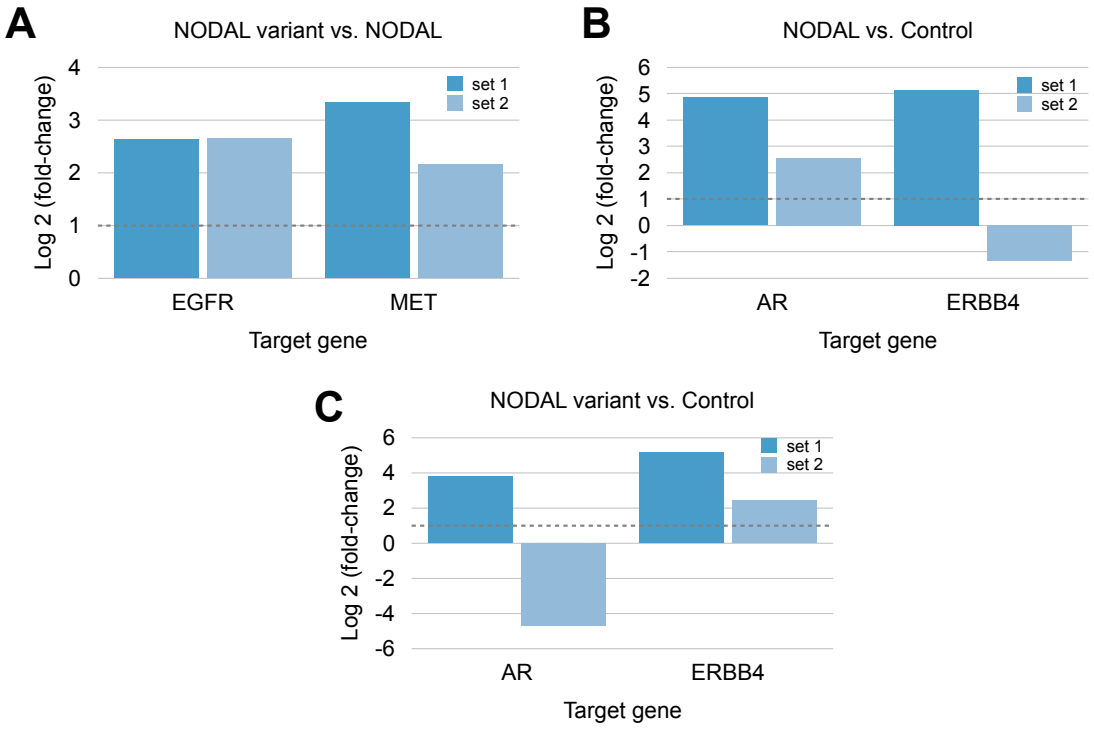


Figure 5.4:Inconsistent expression of select genes related to drug resistance.

Expression of select genes related to drug resistance is not consistently altered by *NODAL* or *NODAL* variant expression. The pairs of cell lines compared are indicated above each chart. Dashed grey lines indicate threshold for differential gene expression (fold change of > 2; log2(fold-change) > 1). Set 1 corresponds to the first set of cell lines used in Figures 5.1-5.3. Set 2 corresponds to a newly generated set of stable cell lines.

Collectively, these results show inconsistent induction of some gene expression changes upon over-expression of *NODAL* isoforms.

These findings underscore a general problem with over-expression models: The stable selection process may provide increased opportunities for genetic and phenotypic drift between stably selected subcultures. This may confound analysis of differences between stable cell lines resulting from expression from the transfected plasmids of interest. This effect may be greater if transfection efficiencies are poor and a small percentage of transfected cells give rise to stable proliferating cells in the presence of antibiotic. In addition, typical transgene silencing and random insertion can result in mosaic gene expression in a population of stably selected cells [34].

The use of an inducible transgene expression system avoids the confounding effects of selection, as cells are stably selected and then divided into control and transgene expressing groups after selection upon the addition of an inducing agent. Furthermore, recent advances in precision genome editing technologies have enabled more efficient gene targeting in human cells. This allows targeted transgene integration at a defined genomic locus to minimize transgene silencing, random integration, and mosaic expression.

Since *NODAL* function has been well characterized in breast cancer, we began with human breast cancer cell lines to develop robust models using genome editing. Using previously designed TALENs targeting the AAVS1 safe harbour locus within the PPP1R12C gene locus, we developed T47D breast cancer cells with targeted integration of inducible *NODAL* variant open reading frame (Figure 5.5). First, the donor plasmid used in [35] with TET-ON driven by a CAG promoter and EGFP under the control of an inducible promoter was modified to contain a LacZ insert for blue/white colony screening flanked by Esp3I typeIIS restriction enzyme sites for one-step cloning (Figure 5.5A). This plasmid was used to clone either constitutive *NODAL* or the *NODAL* variant open reading frames with MYC tags at the N-terminal end of the mature peptide. We next developed a ddPCR assay to screen for successfully targeted AAVS1 loci with integrated donor plasmid. This assay used a forward primer specific for target gDNA sequence

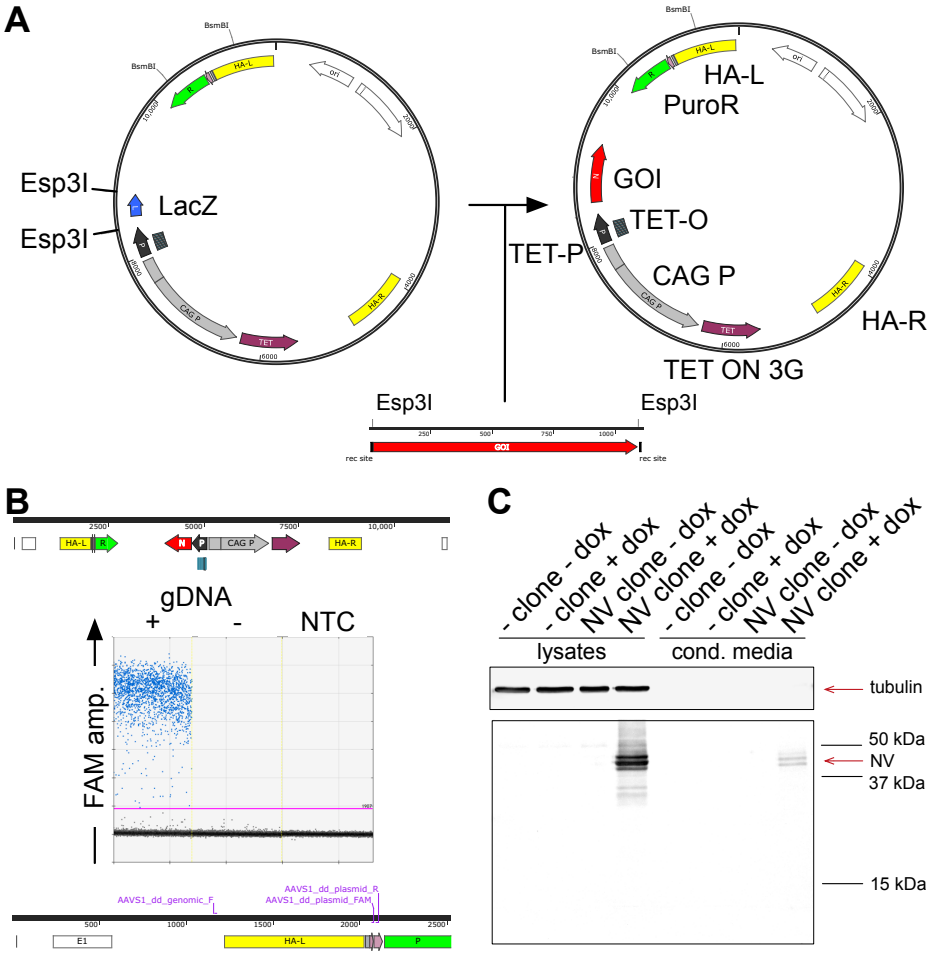


Figure 5.5: Overview of AAVS1-safe harbour targeted transgene expression mediated by precision genome editing in T47D breast cancer cells.
A) A newly constructed AAVS1 donor plasmid contains a LacZα insert for blue/white colony screening flanked by Esp3I sites to facilitate one-step cloning. Insertion of a gene of interest (middle; red) results in a transfection-ready targeting construct. “GOI” = gene of interest. “TET-P” = tetracycline-responsive promoter. “TET-O” = tetracycline-responsive operator. “CAG P” = chimeric constitutive promoter. “TET ON 3G” = protein binding TET operator in the presence of doxycycline. “HA-R” = homology arm right. “HA-L” = homology arm left. “PuroR” = puromycin resistance gene. B) Top: integrated donor plasmid from homology arm left to homology arm right in the context of the AAVS1 genomic locus. White boxes indicate PPP1R12C exons 1 (left) and 2 (right). Middle: ddPCR screening assay for clones positive (+) and negative (-) for AAVS1 integration. “NTC” = no template control. Bottom: position of primers and probes used in ddPCR screening assay. C) Induction of *NODAL* variant expression in clones positive (+) for AAVS1 integration. “dox” = doxycycline. “NV” = *NODAL* variant.

outside of the plasmid homology arm, and a reverse primer and fluorescent probe specific to plasmid sequence not endogenously found in gDNA (Figure 5.5B). When paired with an assay for copy number analysis of the AAVS1 locus outside of the plasmid homology arms, our assay can be used to determine the proportion of AAVS1 loci successfully targeted in a mixed or clonal population of cells. We report successful generation of T47D breast cancer cells with stable donor integration and doxycycline-inducible *NODAL* variant expression. After treatment with doxycycline, *NODAL* variant expression was detectable in both cell lysates and conditioned media (Figure 5.5C). Several bands were evident, likely resulting from alternative N-glycosylation as reported in Chapter 4. However, mature peptide was not readily detected in the conditioned media of T47D cells as it was in HEK 293 cells.

As a complement to over-expression studies, disruption of endogenous gene function is common. Such disruption has also greatly benefited from recent advances in precision genome editing technologies. Instead of relying on variably efficient post-transcriptional inhibition of gene expression using processes such as RNA interference, the induction of mutations can result in stable missense gene expression and complete endogenous functional knockout of a gene of interest.

Droplet digital PCR assays were next developed for specific and sensitive screening of precision-nuclease treated cells for mutations desirable for gene knockout (Figure 5.6). These duplexed assays consisted of forward and reverse primers amplifying the target cut site, a reference probe designed to bind away from the target cut site, and a “drop-off” probe designed to bind the unmutated target cut site (Figure 5.6A). Droplets containing unmutated wild type target gDNA are positive for signal from both probes, while droplets containing mutated target gDNA are positive for reference probe signal but negative for drop-off probe signal, as the latter probe can no longer bind (Figure 5.6A). Assays for two targets in constitutive *NODAL* exon 1 and another target at the *NODAL* cassette alternative exon 5’ splice donor site were designed to test the efficiency of engineered TALEN proteins targeting these loci (Figures 5.7-5.9). TALENs were designed with the NH RVD to target G bases within high G-C content target loci according to guidelines from [36, 37]. For target 1 in constitutive exon 1 with 48% G bases (Figure 5.7A), a

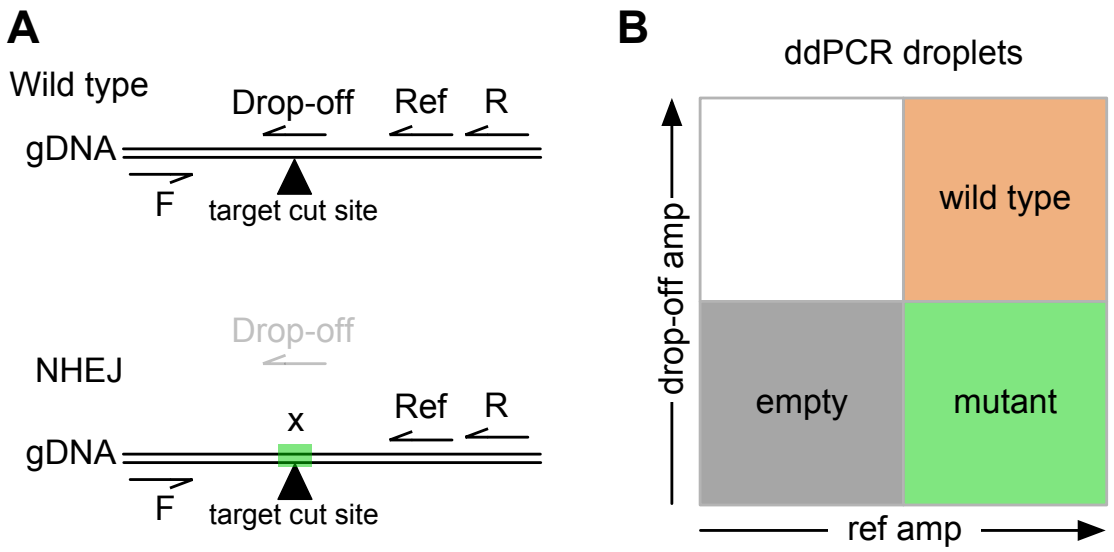


Figure 5.6: Development of ddPCR assays to screen for mutations resulting from non-homologous end joining (NHEJ).
A) General ddPCR assay strategy indicating primer and probe binding sites relative to a target cut site of interest. B) Schematic of 2D ddPCR droplet results for droplets containing mutated or wild type targets. “ref amp” = reference probe amplitude. “drop-off amp” = drop-off probe amplitude.

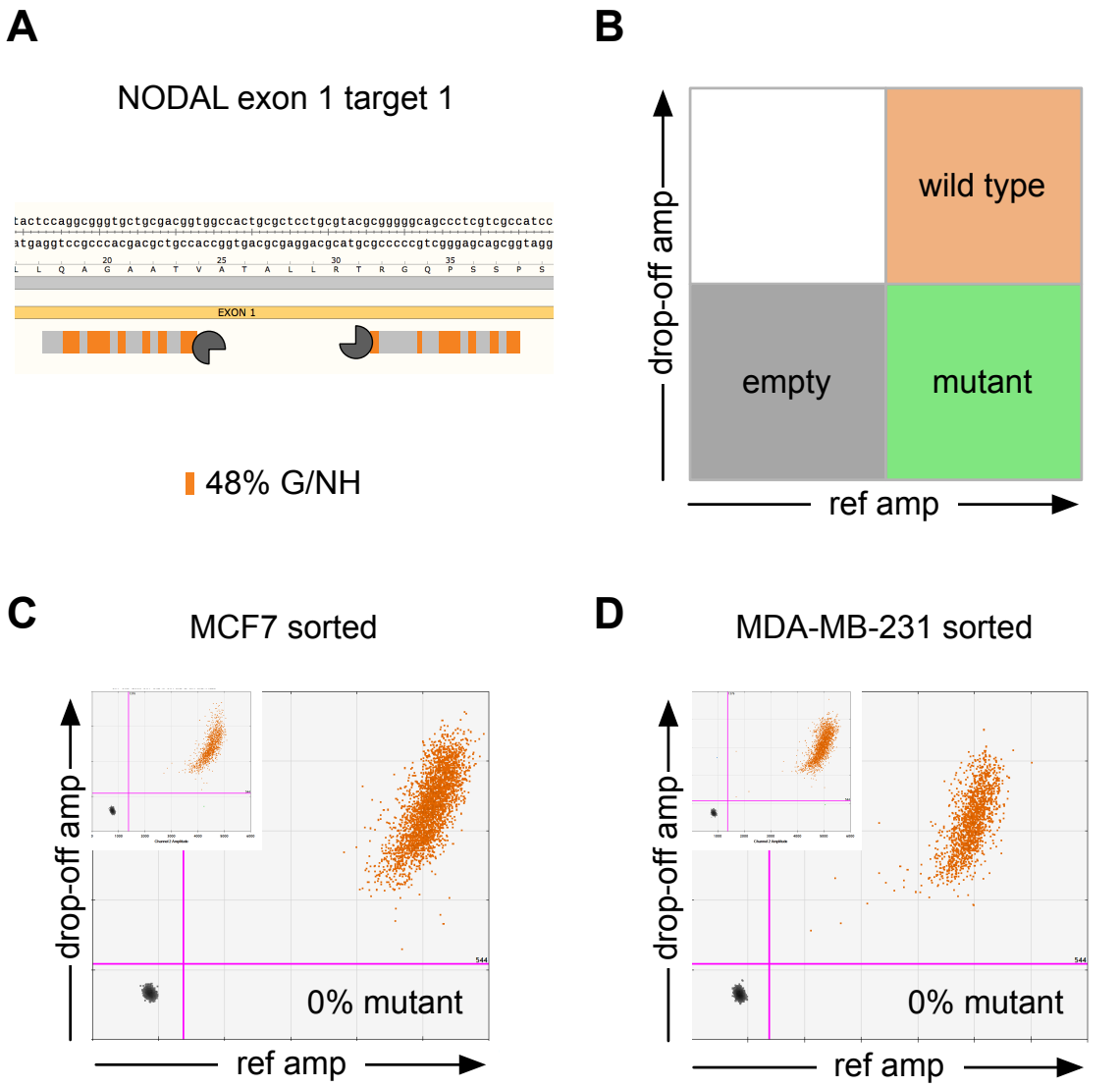


Figure 5.7: TALENs constructed with the NH RVD did not induce mutations at target 1 of constitutive *NODAL* exon 1.
A) Sequence of target 1. TALENs are shown directly under their binding sites. Grey segments indicate HD, NI, or NG RVDs. Orange segments indicate NH RVDs. B) Schematic of 2D ddPCR droplet results for droplets containing mutated or wild type targets. “ref amp” = reference probe amplitude. “drop-off amp” = drop-off probe amplitude. C) No target mutations were detected in MCF7 cells sorted to enrich for TALEN-transfected cells. Droplets for untreated control cells are inset. D) No target mutations were detected in MDA-MB-231 cells sorted to enrich for TALEN-transfected cells. Droplets for untreated control cells are inset.

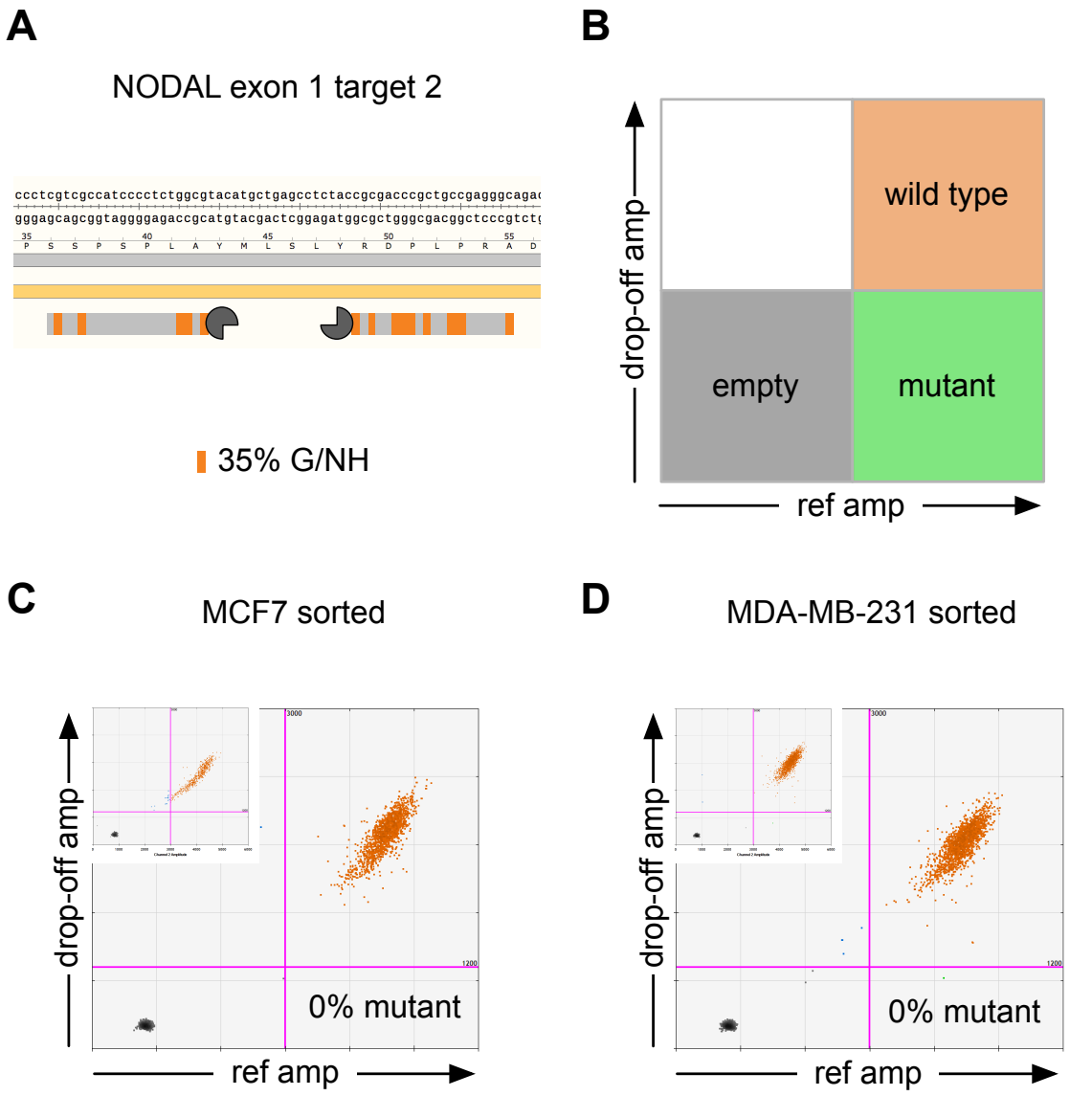


Figure 5.8: TALENs constructed with the NH RVD did not induce mutations at target 2 of constitutive *NODAL* exon 1.
A) Sequence of target 2. TALENs are shown directly under their binding sites. Grey segments indicate HD, NI, or NG RVDs. Orange segments indicate NH RVDs. B) Schematic of 2D ddPCR droplet results for droplets containing mutated or wild type targets. “ref amp” = reference probe amplitude. “drop-off amp” = drop-off probe amplitude. C) No target mutations were detected in MCF7 cells sorted to enrich for TALEN-transfected cells. Droplets for untreated control cells are inset. D) No target mutations were detected in MDA-MB-231 cells sorted to enrich for TALEN-transfected cells. Droplets for untreated control cells are inset.

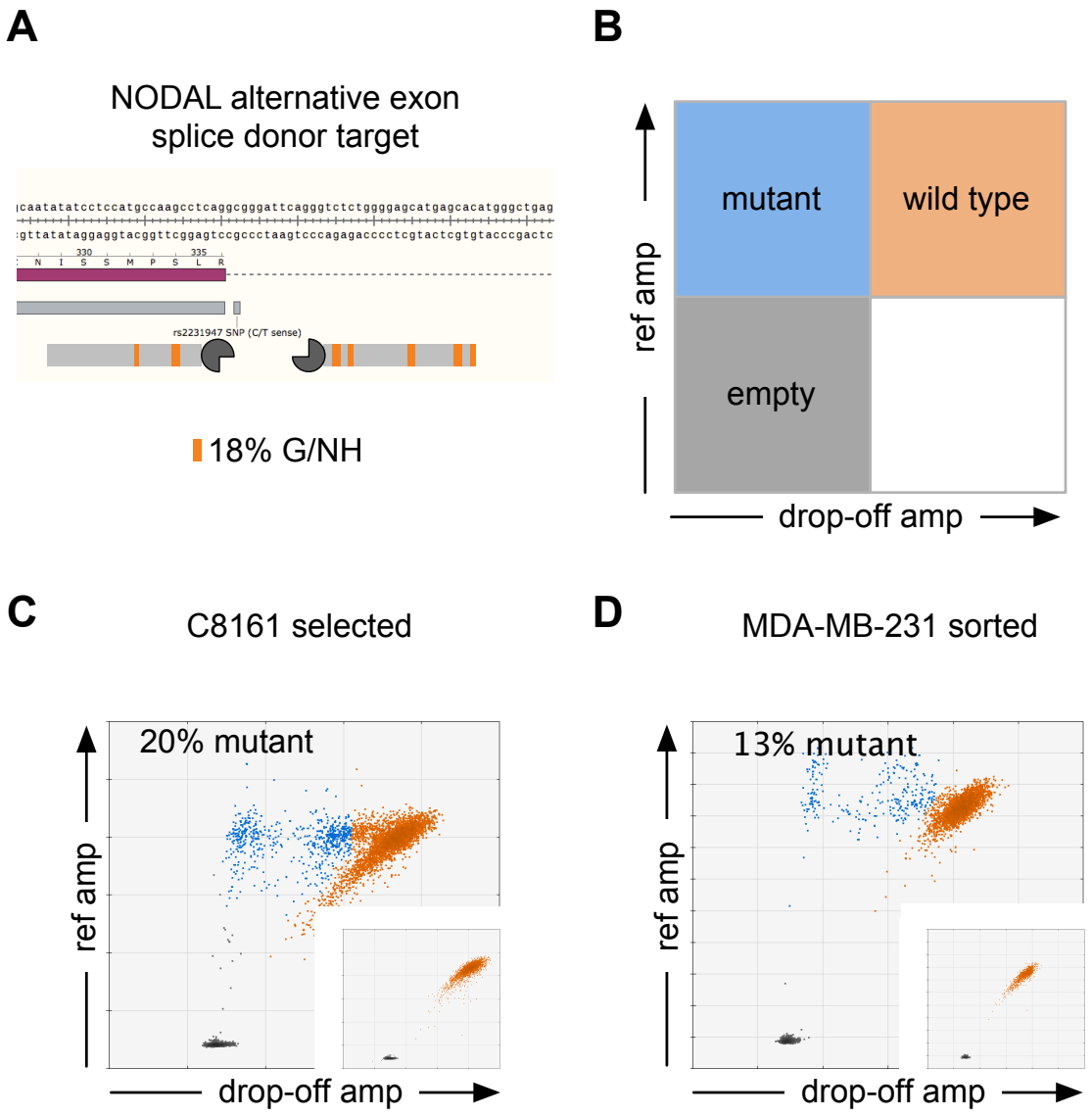


Figure 5.9: TALENs constructed with the NH RVD induce mutations at a *NODAL* alternative exon splice donor target with low guanine content.
A) Sequence of alternative splice donor site target. TALENs are shown directly under their binding sites. Grey segments indicate HD, NI, or NG RVDs. Orange segments indicate NH RVDs. B) Schematic of 2D ddPCR droplet results for droplets containing mutated or wild type targets. “ref amp” = reference probe amplitude. “drop-off amp” = drop-off probe amplitude. C) 20% of target alleles were mutated in C8161 melanoma cells selected to enrich for TALEN-transfected cells. Droplets for untreated control cells are inset. D) 13% of target alleles were mutated in MDA-MB-231 cells sorted to enrich for TALEN-transfected cells. Droplets for untreated control cells are inset.

ddPCR assay did not detect any mutations in TALEN-transfected MCF7 (Figure 5.7C) or MDA MB 231 (Figure 5.7D) breast cancer cells, after fluorescence activated cell sorting (FACS) to enrich for highly transfected cells. For target 2 in constitutive exon 1 with 35% G bases (Figure 5.8A), virtually no mutations were detected (Figure 5.8C-D). For the alternative exon target with only 18% G bases (Figure 5.9A), between 13% and 20% of target gDNA was mutated, indicative of successful genome editing (Figure 5.9C-D). That desired mutations were detected in only targets with low G content prompted us to directly assess the performance of the NH RVD relative to the previously-developed but less G-specific NN RVD [37, 38] for a target with high G content. For an example target with 43% G in exon 1 of the *SFRP1* gene (Figure 5.10A), TALENs with NH RVDs induced mutations at an average frequency of only 1.5%, while TALENs with NN RVDs designed to the same target were an average of 17-fold more efficient, inducing mutations at a frequency of 25% (P=0.0084; Figure 5.10B).

Since the NN RVD lacks high specificity for G bases, and constitutive exon 1 of human *NODAL* has a high GC content, we next explored the CRISPR/Cas9 precision genome editing system for functional knockout of *NODAL*. First, an “all-in-one” plasmid coding for both Cas9 and an associated guide RNA was modified to contain a LacZ insert for blue/white colony screening flanked by Esp3I typeIIS restriction enzyme sites for one-step cloning (Figure 5.11). A CRISPR for target 1 in constitutive *NODAL* exon 1 (Figure 5.12A) induced mutations in 28% of target gDNA (Figure 5.12C-D). A CRISPR for target 2 (Figure 5.13A) induced mutations in 11% of target gDNA (Figure 5.13C-D). Single cell-derived clones were then generated and screened for mutations using ddPCR assays followed by validation with target cloning and Sanger sequencing. Examples of clones with only wild type target alleles, both wild type and mutated target alleles, and only mutated target alleles detected are shown in Figure 5.14. For simplicity, such samples will be referred to as wild type, mono-allele mutation, and bi-allele mutation respectively, although it is possible that *NODAL* is not present at a normal copy number of two in the karyotypically abnormal cancer cell lines used.

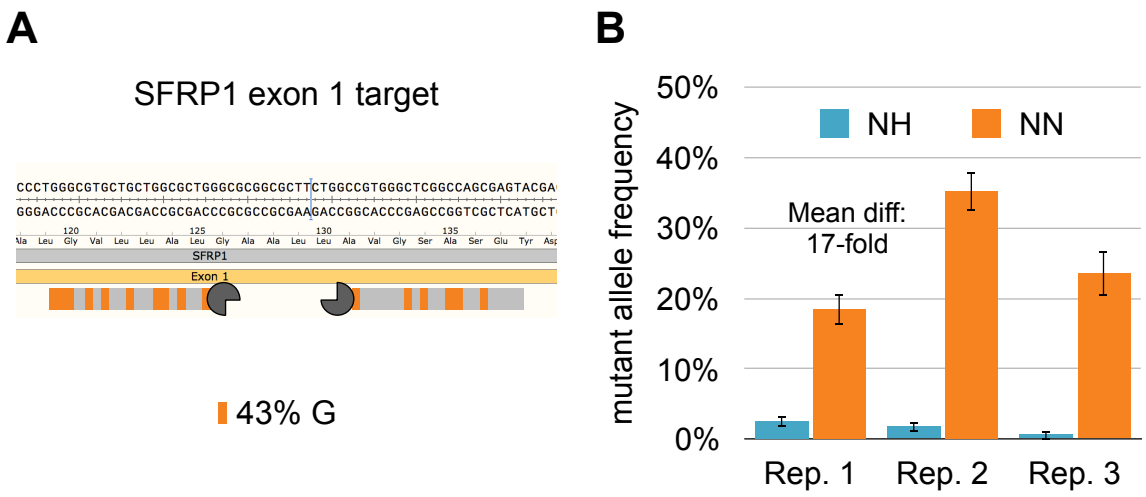


Figure 5.10: TALENs constructed with the less-specific NN RVD to target guanines results in higher mutation efficiencies relative to TALENs with NH RVDs targeting the same locus.

A) Sequence of *SFRP1* exon 1 test target site. TALENs are shown directly under their binding sites. Grey segments indicate HD, NI, or NG RVDs. Orange segments indicate NH/NN RVDs to target guanine (G). Error bars indicate Poisson 95% confidence intervals for each ddPCR assay. Three independent biological replicates ("Rep.") are shown. TALENs with NN RVDs consistently outperformed those with NH RVDs, by a mean difference of 17-fold.

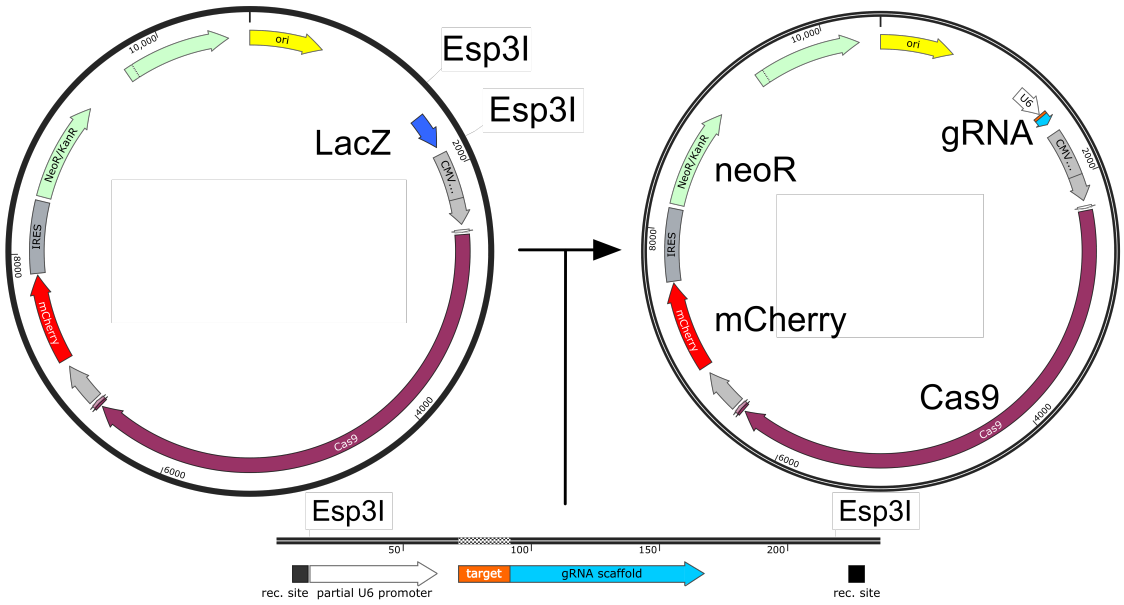


Figure 5.11: Modified CRISPR "all-in-one" plasmid for one step cloning and blue/white screening.

Left: Modified plasmid containing ESp3I sites and LacZ insert. Middle: example of a double stranded DNA insert containing gRNA sequence for target of interest. Right: Example of final plasmid containing desired gRNA sequence. "gRNA" = guide RNA. "neoR" = neomycin resistance gene.

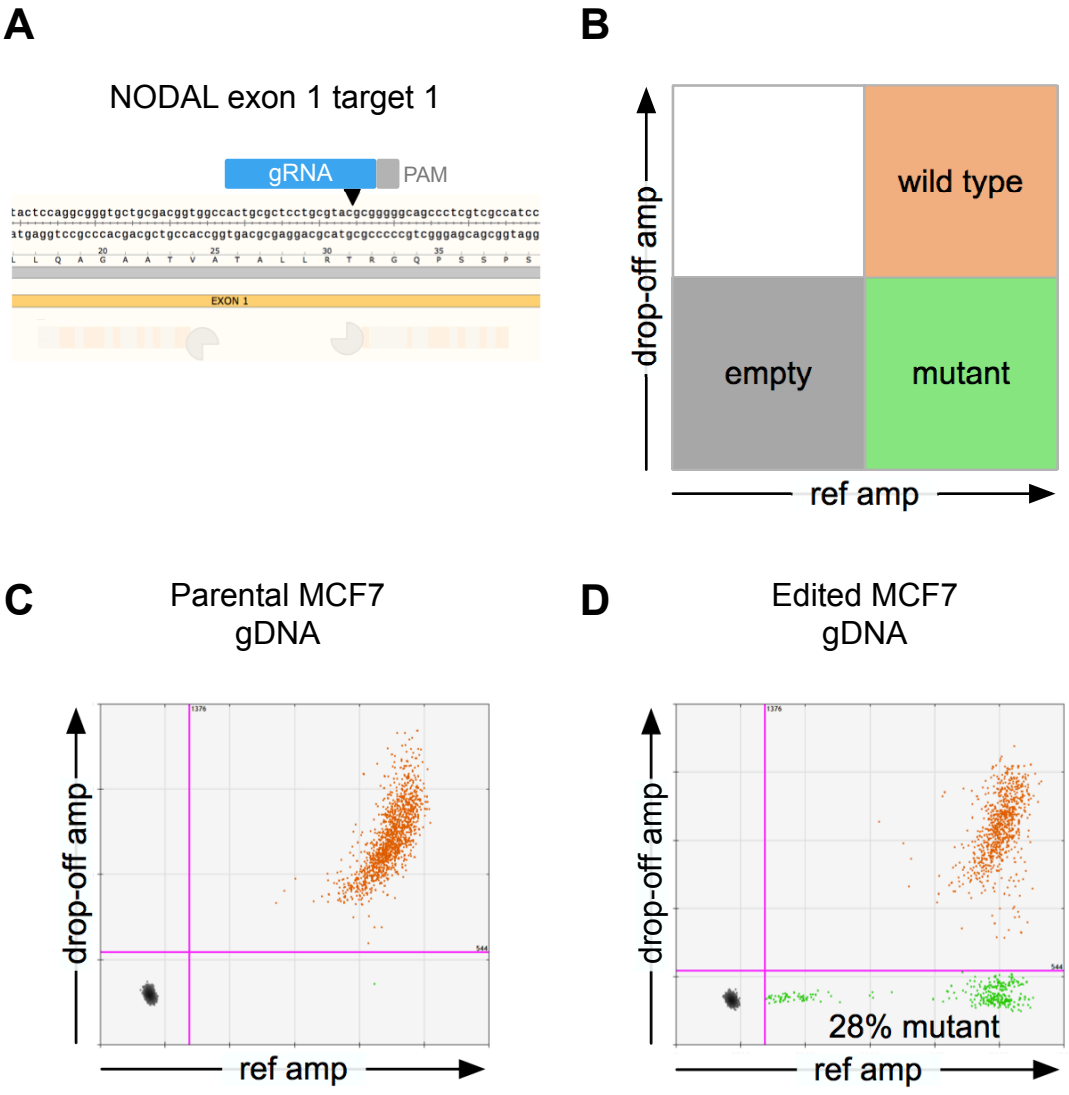


Figure 5.12: A CRISPR gRNA to target 1 of *NODAL* constitutive exon 1 induced mutations in MCF7 cells.
A) Sequence of target 1. The gRNA is shown directly above its binding site. Black arrow indicates expected site of double strand break. TALENs for target 1 from Figure 5.7 are shown (faded) for context. “gRNA” = guide RNA. “PAM” = protospacer-adjacent motif. B) Schematic of 2D ddPCR results for droplets containing mutated or wild type targets. “ref amp” = reference probe amplitude. “drop-off amp” = drop-off probe amplitude. C) Untreated parental MCF7 cells show virtually no mutant droplets. D) 28% of target alleles were mutated in MCF7 cells enriched for CRISPR-transfected cells.

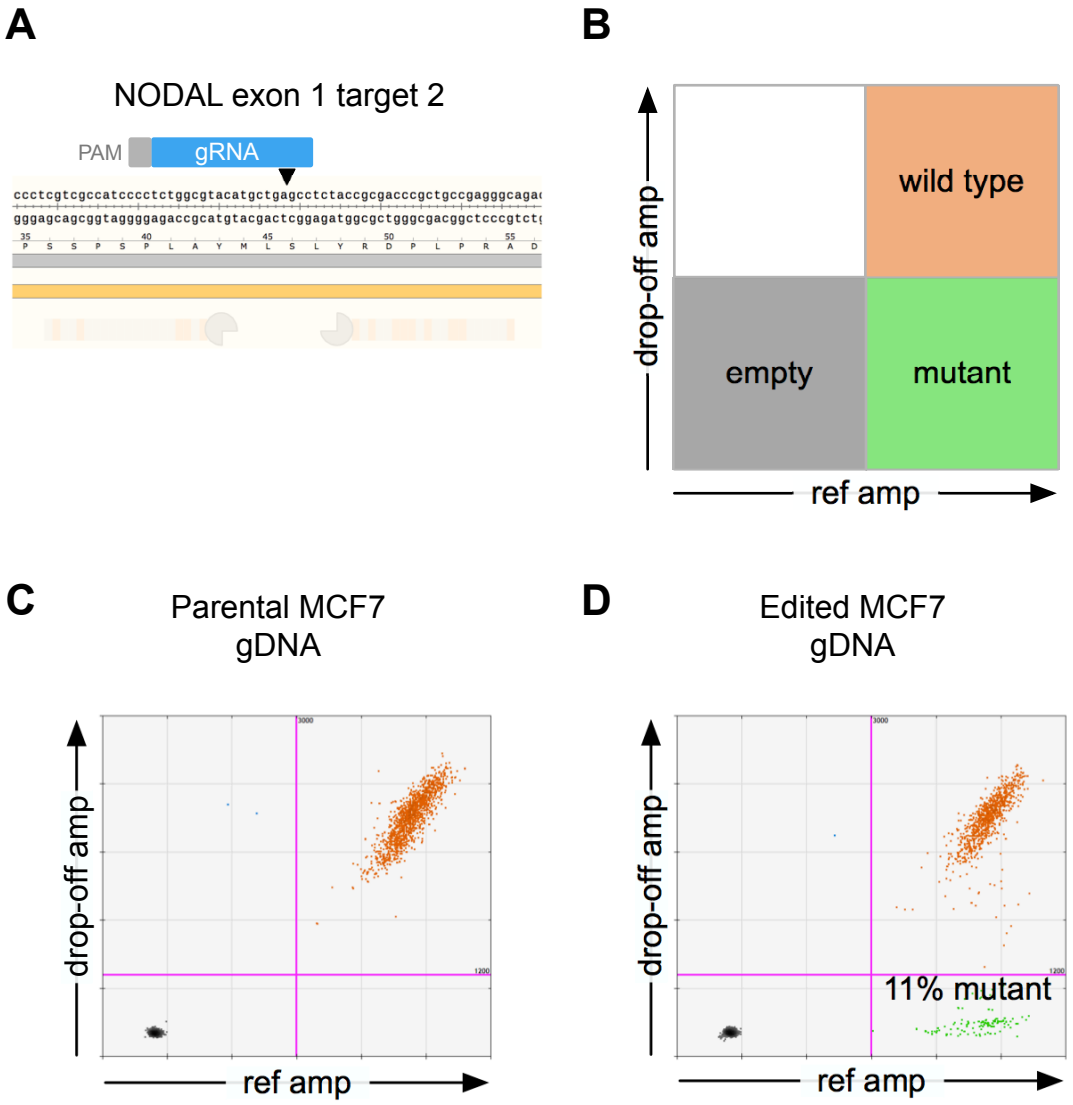


Figure 5.13: A CRISPR gRNA to target 2 of *NODAL* constitutive exon 1 induced mutations in MCF7 cells.
A) Sequence of target 2. The gRNA is shown directly above its binding site. Black arrow indicates expected site of double strand break. TALENs for target 2 from Figure 5.8 are shown (faded) for context. “gRNA” = guide RNA. “PAM” = protospacer-adjacent motif. B) Schematic of 2D ddPCR results for droplets containing mutated or wild type targets. “ref amp” = reference probe amplitude. “drop-off amp” = drop-off probe amplitude. C) Untreated parental MCF7 cells show no mutant droplets. D) 11% of target alleles were mutated in MCF7 cells enriched for CRISPR-transfected cells.

These clones were next used to compare the performance of developed ddPCR assays to mismatch nuclease assays for detection of precision nuclease-induced mutations. Virtually all droplets were classified as wild type for clones with only wild type target alleles. Both wild type and mutant droplet clusters were detected for clones with mono-allelic target mutation, while virtually all droplets were classified as mutant for clones with bi-allelic target mutation (Figure 5.14A). Thus, our ddPCR assays could very easily distinguish clones with partially mutated target alleles from those with fully mutated target alleles. In contrast, a mismatch nuclease assay for the same target 2 samples did not distinguish between partially mutated and fully mutated samples (Figure 5.14B).

We next assessed the quantitative performance of ddPCR assays relative to mismatch nuclease assays. We spiked in different amounts of genomic DNA (gDNA) from the fully mutated target 2 clone into a high concentration of non-mutated wild type gDNA from the unmutated clone. This allowed us to create samples analogous to a small number of mutated cells in a larger background of non-mutated cells, while maintaining the natural complexity, concentration, and purity of a typical gDNA sample.

With respect to sensitivity, even though it required concentrated and purified PCR product as input, the mismatch nuclease assay performed very poorly. In our assay, 0.6% mutant DNA (2.5 ng mutant in 400 ng total PCR product) was difficult to distinguish from background noise (Figure 5.15A). The absolute sensitivity of this assay was very poor (0.6% is 2.5 ng of mutant PCR product, which is approximately 4×10^9 copies of DNA), despite the large amount of input gDNA required to generate sufficient PCR product. In our ddPCR assays, we were able to successfully detect a minimum of between 20 pg and 156 pg of mutant gDNA (not purified PCR product) in a high background of 100 ng of wild type gDNA for our three targets (Figure 5.15B). We did not test below 20 pg as this amount of gDNA was expected to contain between only 1 and 4 copies of target DNA (see methods) and thus served as a practical lower limit. In terms of relative abundance, 20 pg of mutant DNA in 100 ng of wild type DNA is a mutant frequency of 0.02%.

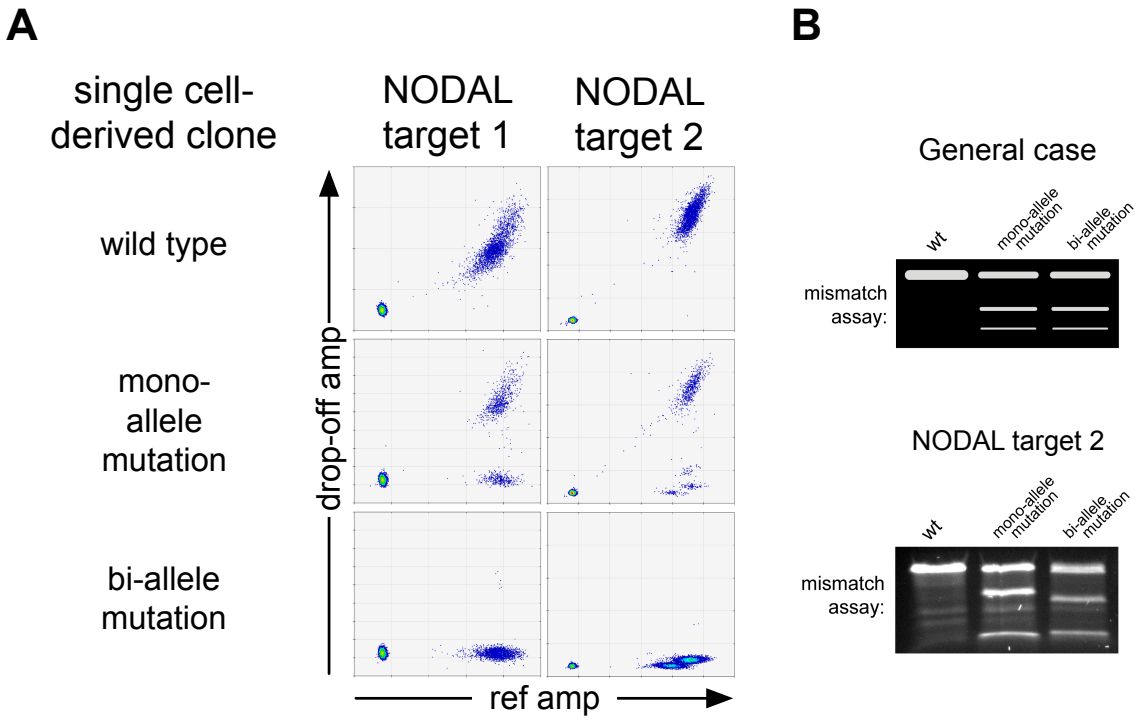


Figure 5.14: For genome editing of *NODAL*, a ddPCR mutation assay outperforms a mismatch nuclease assay in its ability to distinguish mono-allelic mutations from mutations that affect all target alleles.
A) Droplet results in heat map form for ddPCR assays for *NODAL* target 1 and target 2 for mutations with different target mutation profiles. “ref amp” = reference probe amplitude. “drop-off amp” = drop-off probe amplitude. B) Top: expected results of a mismatch nuclease assay for samples with different target mutation profiles. Bottom: actual results of a mismatch nuclease assay for *NODAL* target 2.

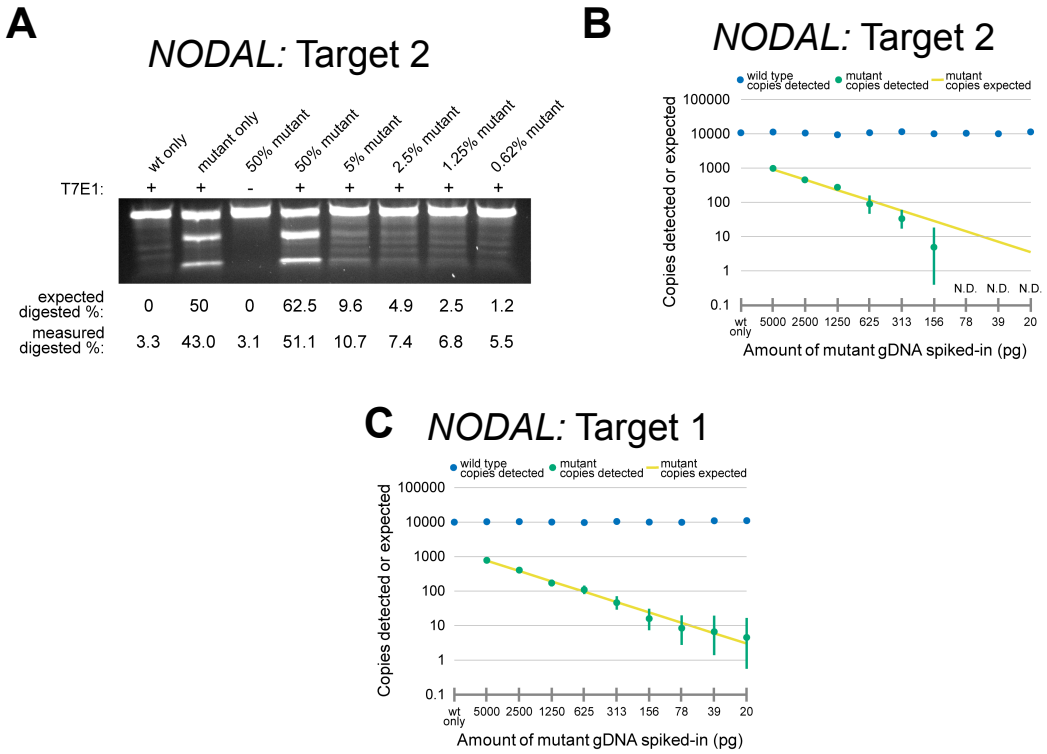


Figure 5.15: ddPCR mutation assays were more accurate, sensitive, and specific in detecting *NODAL* target mutations than a corresponding mismatch nuclease assay

A) A ddPCR assay reliably detected gDNA with mutations at target 1 in a high background of wild type gDNA. B) A mismatch nuclease assay was not particularly accurate or sensitive in detection of mutant gDNA. “T7E1-” sample is a negative control with no nuclease. C) A ddPCR assay reliably detected gDNA with mutations at target 2 for concentrations above 156 ng in a high background of wild type gDNA. “N.D” = not detected.

We were also interested in comparing the accuracy of mismatch assays and ddPCR assays. Our mismatch nuclease assay was not very accurate in its quantification of mutant PCR product at any of the dilutions tested (Figure 5.15A). In the corresponding ddPCR assay, the 95% confidence intervals for the detected copies of mutant gDNA generally encompassed the expected number of mutant copies from 313 pg to 5 ng of mutant genomic DNA (Figure 5.15B). In the ddPCR assay for target 1, quantification of mutant gDNA was accurate from 20 pg to 5 ng (Figure 5.15C). For both assays, the amount of wild type gDNA detected by the ddPCR assay remained stable across samples and was not affected by the amount of mutant gDNA loaded as there was no significant correlation between mutant gDNA loaded and copies of wild type detected (coefficient of determination; $R^2 = 0.014$ for *NODAL* target 1, and $R^2 = 0.045$ for *NODAL* target 2). These data demonstrate that these assays are capable of accurate detection of extremely rare mutations in 100 ng of gDNA.

5.3 Discussion

In studying the impact of *NODAL* over-expression on ovarian cancer cell resistance to chemotherapy, we found the two *NODAL* isoforms to have similar yet distinct impacts. The constitutive *NODAL* isoform conferred more robust resistance to cisplatin relative to the *NODAL* variant. In the absence of chemotherapy exposure, perhaps surprisingly, *both NODAL* isoforms induced remarkably similar changes in the expression of genes related to drug resistance, and promoted clonogenic growth to similar extents. Given the divergence in C-terminal peptide sequence between constitutive *NODAL* and the *NODAL* variant, and the *NODAL* variant's inability to induce expression of targets of canonical *NODAL* signalling in an embryonic system, it is unlikely that the *NODAL* variant is affecting cancer cell plasticity and chemotherapy resistance via a canonical *NODAL* signalling response. Therefore, there are several possibilities for the observed phenotypes resulting from *NODAL* variant over-expression: Despite lacking canonical function in a regulated embryonic system, the *NODAL* variant may be able to transduce a diminished canonical *NODAL* signal in some contexts. It is also possible that both *NODAL* isoforms can engage non-canonical signalling pathways possibly involving currently unidentified receptor complexes or hetero-dimerization with other related ligands. Lastly, the NODAL

pro-peptide (which is common between both *NODAL* splice variants) may also have pro-tumourigenic function independent of the mature peptides.

Regardless of the mechanisms by which the *NODAL* isoforms induced changes in ovarian cancer cells, the *NODAL* variant seemed to have a similar yet more limited impact on pro-tumourigenic phenotypes of A2780S ovarian cancer cells. However, the dramatic changes in gene expression of several genes related to drug resistance in control cells during stable selection highlights a drawback of over-expression models that is particularly problematic in cancer cell lines. Phenotypic drift resulting from bottlenecking of genetically and epigenetically heterogeneous cells introduces a confounding variable to studying changes truly induced by the activity of a gene product of interest. Even after selection, variability in transgene integration and expression within a stable population of cells can also result.

Since precision genome editing potentiates attractive alternative models to both conventional over-expression studies such as those performed for *NODAL* here, as well as variably efficient post-transcriptional knockdown approaches, we sought to develop more robust models to study *NODAL* biology in cancer systems using genome editing. We successfully generated cancer cell lines with an inducible *NODAL* variant expression construct integrated at the AAVS1 safe harbour locus, and with reading frame-altering mutations in constitutive exon 1 for functional *NODAL* knockout. These models are currently being used to evaluate the performance of NODAL protein detection assays and as robust models to explore *NODAL* function in cancer systems.

Streamlined quantitative screening of nuclease-edited cells is imperative for genome editing to reach its full research potential. Thus, we developed ddPCR mutation detection assays, using *NODAL*-edited cells to demonstrate their utility. These assays can be easily adapted to any desired target, and will be of value for the many research fields utilizing precision genome editing techniques. Guidelines for ddPCR mutation screening assay design are included in Appendix C.

Beyond detailing ddPCR mutation screening assays, we directly tested them against the widely utilized mismatch nuclease assay. We demonstrated that ddPCR assays are more

specific, accurate, and sensitive. They also offer practical advantages: First, only a small amount of gDNA (as little as 5 ng total gDNA) is required for analysis. In contrast, a relatively large amount of input DNA (e.g. 500 ng) is generally required to generate 400 ng of purified target PCR product required for mismatch nuclease assays. Second, these assays easily discriminate between single-cell derived clones with mono-allelic mutations and those with both alleles successfully mutated. Third, ddPCR assays are more versatile in that they can more easily avoid false-positive mutation calls due to pre-existing mutations or SNPs outside of the nuclease target site. Together, these characteristics translate to a much more rapid and efficient workflow for the user.

Importantly, all of the samples used in this study were genomic DNA preparations and not highly purified PCR products, synthetic oligos, or gene fragments. This allowed us to test the practical utility of these assays in prototypical samples. The theoretical limit of sensitivity for any mutation detection assay is detection of a single mutated molecule in a high background of wild type molecules. One of our two assays was able to distinguish only 20 pg of mutant DNA from 100 ng of wild type DNA (0.02%). Given that a typical diploid human cell is estimated to contain about 6 pg of gDNA and we were using karyotypically abnormal cancer cell lines, it is likely that 20 pg is very close to the biological limit of detection, representing all the alleles from a single cell.

Beyond showcasing their utility, we also used ddPCR assays to demonstrate poor genome editing performance of TALENs built with the NH RVD for guanine-rich targets. Thus, it may not be advisable to maximize GC target content for TALENs using the NH RVD. Indeed, widely followed design guidelines [37] available as options in TALEN design software and assembly kits [36, 39] suggest to target loci with at least 25% C+G and avoid stretches of 6 or more A+T. This recommendation was initially made based on the identification of NN (targeting G) and HD (targeting C) as “strong binders” that stabilized TALEN-DNA binding [37]. However, since these recommendations were published, NH has become widely adopted as the G-targeting RVD of choice due to increased specificity over NN [37, 38, 40]. Unfortunately, the strength of NH binding appears to be context dependent and has been characterized as an “intermediate binder.” Specifically, unlike NN, using NH to target G did not result in any TALEN activity for an

A+ T rich 11 bp target lacking any C nucleotides [37]. If the design guidelines of >25% C+G are extended to “maximize C+G” and the NH RVD is employed, we hypothesized that TALEN activity may suffer. Indeed, this was the case for our G-rich *NODAL* exon 1 targets.

We have demonstrated that ddPCR mutation detection assays have great utility and offer several benefits over conventional mutation screening methods. They are ideal for rapid genome editing workflows as they require very little sample genomic DNA, and the same assay can be used for screening bulk populations and single cell-derived clones. These assays will undoubtedly continue to increase in popularity and contribute to rapid and quantitative genome editing workflows.

Upstream of screening for successfully edited targets, we also cloned new versions of both an AAVS1 donor plasmid and an “all-in-one” plasmid for CRISPR/Cas9 editing. These plasmids allow for simplified and more efficient cloning for any desired target, further streamlining genome editing workflows. Guidelines for cloning a desired target gRNA into the “all-in-one” CRISPR plasmid are detailed in Appendix C.

The functional *NODAL* knockout and inducible over-expression cell lines generated and validated with these newly developed genome editing tools are currently being used to further understand *NODAL* biology as it pertains to cancer phenotypes. Specifically, the *NODAL* knockout cell lines are being used to validate the specificity of NODAL antibodies and to validate *NODAL* expression at the protein level in cancer cell lines. As the efficiency of precision genome editing continues to improve, this general approach can be used to experimentally manipulate specific elements of *NODAL* such as SNPs, splice sites, polyadenylation signals, and sites of post-translational modification. Since genome editing results in stable and heritable mutations, these modifications will potentiate robust modelling of *endogenous NODAL* expression and function.

5.4 Methods

5.4.1 Cell culture

A2780S cells were cultured in DMEM/ FF12 media supplemented with 10% fetal bovine serum (FBS; Thermo Fisher; Waltham, Massachusetts, USA). T47D, MCF7, and MDA-MB-231 breast cancer cells, and C8161 melanoma cells [41], were cultured in RPMI supplemented with 10% FBS (Thermo Fisher). All cells were cultured at 37°C with 5% $C0_2$ supplementation.

5.4.2 MTT assays

Cells were seeded in 96 well plates at a density of 5,000 cells/well in 100 µl of complete medium and exposed to varying concentrations of carboplatin (3.1-200 µM) for 72 hours. MTT reagent (10 µL of 5 mg/mL in PBS) was added to each well for 4 hours. After 4 hours, the resultant formazan crystals were dissolved in 100 µL of solubilization solution and the absorbance at 570 nm was measured with a reference wavelength of >650 nm. All experiments were performed in triplicate (3 technical and 3 biological replicates), and the relative cellular viability (%) was expressed as a percentage relative to untreated control cells.

5.4.3 Clonogenic growth assays

A modified version of the protocol from [42] was used. Cells were seeded in 6-well plates at the following densities: 50 cells/ well for no carboplatin treatment, 500 cells /well for 3.1 µM carboplatin, 1,000 cells/ well for 6.3 µM, 1,500 cells/well for 12.5 µM, 2,000 cells/well for 25 µM, and 2,500 cells/well for 50 µM). Six hours post-seeding, cells were treated with the appropriate carboplatin dose for 24 hours. The medium with carboplatin was then replaced with fresh medium and cells were allowed to grow for 7-10 days in the absence of carboplatin treatment. The colonies formed were then gently washed with PBS, fixed with methanol/acetic acid (3:1) solution stained with crystal violet (0.5% in methanol). Colonies of ≥ 50 cells were counted and the viability was calculated using equations: Plating efficiency (PE)= count of colonies formed in control wells/number of cells seeded in control wells; Relative clonogenic viability (%)= count of colonies formed in treated wells/ number of cells seeded in these wells / PE x 100.

5.4.4 Colony formation assays

The soft agar colony formation assay was used to assess cellular anchorage-independent growth *in vitro.* Cells were suspended in 2x medium containing 0.7% low melting agarose (1:1), and plated onto solidified 1% agarose containing 2x medium (1:1) in 6-well culture dishes at a density of 2,000 cells per well. Cells were incubated for 2 weeks (medium was changed every 2-3 days). The colonies formed were then washed with PBS and stained with crystal violet (0.5% in methanol). Number of colonies formed was expressed as colony formation efficiency relative to control (A2780s GFP) cells.

5.4.5 PCR arrays

Total RNA was isolated from cultured cells using the PerfectPure RNA Cultured Cell Kit (5-Prime; Hilden, Germany) including on-column DNase treatment and quantified with the Epoch plate reader (Biotek; Winooski, Vermont, USA). Reverse transcription of RNA was performed using the RT² First Strand cDNA Kit (SABiosciences/ Qiagen; Hilden, Germany) according to the manufacturer's instructions. The human "Cancer Drug Resistance PCR Array" RT^2 Profiler PCR Array (SA Biosciences/ Qiagen) was used for SYBR green real time PCR detection of genes related to cancer cell drug resistance. Plates were cycled according to manufacturer's instructions using the CFX 96 real time PCR system and results were analyzed with CFX manager (Bio-Rad; Hercules, California, USA). Melt curve analysis was used to exclude samples with low melt peaks and inconsistent melt profiles for the same target between samples, indicative of off-target amplification. Expression values were normalized to the mean expression of four endogenous control targets (ACTB, GAPDH, HPRT1, and RPLP0) included in the array, using the ΔΔCt method.

5.4.6 Plasmid cloning

A previously generated AAVS1 donor plasmid and associated TALENs for genome editing were gifts from Su-Chun Zhang (inducible donor plasmid: Addgene; Cambridge, Massachusetts, USA; plasmid # 52343; Left TALEN: # 52341; Right TALEN: # 52342). The donor plasmid was modified using site-directed mutagenesis to introduce Type IIS BsmBI/ Esp3I restriction sites flanking the inducible EGFP open reading frame. Site-

directed mutagenesis was performed with the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent; Santa Clara, California, USA) according to manufacturer's instructions. These sites were used to insert a LacZα cassette for blue/white colony screening. Note that other BsmBI sites are present in the plasmid and these regions need to be verified by sequencing after plasmid construction. The *NODAL* variant open reading frame containing a MYC tag at the N-terminal end of the mature peptide was inserted into this plasmid in place of the BsmBI-flanked LacZα insert for AAVS1 targeting and inducible expression.

5.4.6.1 TALEN plasmids and targets

TALEN targets were designed using the TAL Effector Nucleotide Targeter 2.0 (https://tale-nt.cac.cornell.edu/node/add/talen and [36, 39]), using either NH or NN to target G nucleotides, the Streubel et al. guidelines "on," and the upstream base as "T only." The plasmid kit used for generation of TALENs was a gift from Daniel Voytas and Adam Bogdanove (Addgene kit # 1000000024; https://www.addgene.org/taleffector/goldengatev2/ , and [39]) using either the NN or NH RVD to target G nucleotides. In cases where the most 3' nucleotide was G, NH (and not NN) was always used for the last half repeat. Plasmids pTAL7a and pTALb were gifts from Boris Greber (Addgene plasmid # 48705, [43]) and were used as final destination plasmids.

TALEN target sequences:

NODAL exon 1 target 1 left/ sense:	CCAGGCGGGTGCTGCGACGG
NODAL exon 1 target 1 right/ anti-sense:	GGCGACGAGGGCTGCCCCCG
NODAL exon 1 target 2 left/ sense:	CGTCGCCATCCCCTCTGGCG
NODAL exon 1 target 2 right/ anti-sense:	GCCCTCGGCAGCGGGTCGCG
NODAL alternative exon left/ sense:	ATATCCTCCATGCCAAGCCT
NODAL alternative exon right/ anti-sense:	GTGCTCATGCTCCCCAGAGA
SFRP exon 1 left/ sense:	GGGCGTGCTGCTGGCGCTGG
SFRP exon 1 right/ anti-sense:	ACTCGCTGGCCGAGCCCACG

5.4.6.2 CRISPR plasmids and targets

The all-in-one CRISPR/Cas9 LacZ plasmid was generated from the "scrambled sgRNA control for pCRISPR-CG01" plasmid (Genecopedia; Rockville, Maryland, USA). Two unique BsmBI restriction sites flanking the gRNA sequence were consecutively introduced using the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent). The LacZα fragment was then cloned into the plasmid using BsmBI, replacing the original gRNA. The final plasmid ready for one-step cloning is available from Addgene (https://www.addgene.org/74293/). Guidelines for custom gRNA cloning using this plasmid are provided in Appendix C.

CRISPR gRNAs used:
NODAL exon 1 target 1: ACTGCGCTCCTGCGTACGCG
NODAL exon 1 target 2: GGCTCAGCATGTACGCCAGA

5.4.7 Transfections

Transfections were performed with GeneIn transfection reagent (GlobalStem; Rockville, Maryland, USA) according to manufacturer's instructions. For enrichment of AAVS1 targeted clones, transfected cells were selected with puromycin (Thermo Fisher) at a concentration of 1 μg/mL. TALEN transfected cells were enriched using flow cytometry to collect GFP^+ cells, or selected using puromycin (0.5 μg/mL) and blasticidin (2 μg/mL). For enrichment of CRISPR transfected cells, transfected cultures were either selected with 600-1000 μg/mL Geneticin (Thermo Fisher), or sorted for mCherry+ cells using flow cytometry (Faculty of Medicine and Dentistry Flow Cytometry Facility at the University of Alberta). Single-cell derived clones were generated using either flow cytometry to plate a single cell per well of a 96 well plate, or filtered using a 40 μM filter (Thermo Fisher) and manually plated at a concentration of 0.5 cells/ well.

5.4.8 Genomic DNA isolation

Genomic DNA was isolated using the PureLink Genomic DNA isolation kit (Thermo Fisher) and quantified using the Epoch Microplate Spectrophotometer (BioTek).

5.4.9 Droplet digital PCR assays

Droplet digital PCR assays consisted of the following components, with final concentrations indicated in parentheses: ddPCR SuperMix for Probes (no dUTP) (1x, Bio-Rad), forward primer (900 nM), reverse primer (900 nM), Reference probe (250 nM), NHEJ/drop-off probe (250 nM), restriction enzyme (variable based on assay, 4 units). All primers and probes were designed using Primer3 plus (http://primer3plus.com) and purchased from IDT DNA (Coralville, Iowa, USA). All probes included the ZEN internal quencher and 3' Iowa Black FQ quencher. All ddPCR assays were analyzed using the QX200 droplet reader and Quantasoft software version 1.7.4 (Bio-Rad). Standard ddPCR thermal cycling conditions were used for most assays, with an annealing temperature of 55°C. For *NODAL* exon 1 assays, a "3-step" protocol was used, with an annealing temperature of 56°C and an additional 2 minute extension step at 72°C performed for each cycle. Guidelines for primer and probe design are provided in Appendix C.

Primers and probes used:

NODAL exon 1 forward primer: TTCCTTCTGCACGCC

NODAL exon 1 reference probe:

TGGGCCCTACTCCAGG (/5HEX/TGGGCCCTA/ZEN/CTCCAGG/3IABkFQ/)

NODAL exon 1 target 1 drop-off/ NHEJ probe:

CCGCGTACGCAGGAGC (/56-FAM/CCGCGTACG/ZEN/CAGGAGC/3IABkFQ/)

NODAL exon 1 target 2 drop-off/ NHEJ probe:

CTCAGCATGTACGCCAGAG

(/56-FAM/CTCAGCATG/ZEN/TACGCCAGAG/3IABkFQ/)

NODAL exon 1 reverse primer: TAGGCTGCGGATGATG

NODAL alternative exon forward primer: TTGCAATATATCCTCCATGCCA

NODAL alternative exon reference probe:

AAGCTCTAGTACCCCCAGGGA

(/56-FAM/AAGCTCTAG/ZEN/TACCCCCAGGGA/3IABkFQ/)

NODAL alternative exon drop-off/NHEJ probe:

ACCCTGAATCCCGCCTGAG

(/5HEX/ACCCTGAAT/ZEN/CCCGCCTGAG/3IABkFQ/)

NODAL alternative exon reverse primer: GGTGAGGCTCAGGACAGAT

SFRP1 ddPCR forward primer: CATGGGCATCGGGCG

SFRP1 ddPCR reference probe:

CTGGGCGTGCTGCTGG (/56-FAM/CTGGGCGTG/ZEN/CTGCTGG/3IABkFQ/)

SFRP1 ddPCR drop-off/ NHEJ probe:

CGCGGCGCTTCTGGC (/5HEX/CGCGGCGCT/ZEN/TCTGGC/3IABkFQ/)

SFRP1 ddPCR reverse primer: CGTAGTCGTACTCGCTGG

AAVS 1 integration screen forward primer (genomic): TTGAGCTCTACTGGCTTC

AAVS 1 integration screen reverse primer (plasmid): GCATGTTAGAAGACTTCCTC

AAVS 1 integration screen probe:

TCTCCGCTGCCAGATCTC

(/56-FAM/TCTCCGCTG/ZEN/CCAGATCTC/3IABkFQ/)

5.4.10 Mismatch nuclease assay

For the T7E1 mismatch assay, genomic DNA was PCR amplified using AmpliTaq Gold 360 Master Mix (Thermo Fisher), purified using the PureLink PCR Purification Kit (Thermo Fisher) and quantified using the Epoch Microplate Spectrophotometer (BioTek). 400 ng of purified PCR product was used in an annealing reaction and subsequent T7E1 digestion (New England BioLabs; Whitby, Ontario, Canada) as previously described [44]. Cleavage was visualized by agarose gel electrophoresis and detection using the AlphaImager HP (Bio-techne; Minneapolis, Minnesota, USA). Band intensities were obtained by AlphaView software (Bio-techne). Analysis “bands” were placed so as to completely encompass each visible band. Where bands were difficult to visualize, analysis bands were placed in the same location as adjacent wells to provide an unbiased quantification. All analysis bands for bands of a given size were the same width across all lanes. The detected percent digested was calculated as the sum of the intensities of the digested fragment bands divided by the sum of the intensities of all bands. The expected percent digested was determined by assuming random hybridization of alleles and determining the expected frequency of heteroduplexes.

NODAL mismatch forward primer: TCCCCAGAGGGAGGAAAGG

NODAL mismatch reverse primer: CAGGCTCCGGGATAAGCAAC

5.4.11 Dilution series analysis using ddPCR

For the ddPCR dilution series, negative control (wild-type only) reactions and positive control (mutant only) reactions were used to assign thresholds for all dilution sample wells. The wild type population was quantified by setting all other droplets as FAM-negative and HEX-negative. The NHEJ population was quantified manually using the equation: copies/ 20 µL sample = -ln(1-p) x 20,000 / 0.85. 'p' is the proportion of positive droplets defined as NHEJ droplets/ (NHEJ droplets + empty droplets), and 0.85 nL is the average volume of a droplet as used by QuantaSoft (Bio-rad) [45]. Note that for quantification of NHEJ, wild type droplets are excluded from the calculation, as an indistinguishable subpopulation of wild type droplets will also contain NHEJ targets. The expected number of copies was calculated based on the number of copies detected by ddPCR in 100 ng (as measured by spectrophotometry) of each mutant sample.

5.4.12 Sequencing of single cell-derived clones

All single cell-derived clones used for ddPCR and ongoing functional *NODAL* knockout were validated using Sanger sequencing of the intended nuclease target. PCR products for each target were generated and cloned using the TOPO TA Cloning Kit (Thermo Fisher), minipreped using the Diamed High-Speed Plasmid Mini Prep Kit (Frogga Bio; Toronto, Ontario, Canada), and Sanger sequenced by the Molecular Biology Service Unit, University of Alberta. Several clones were sequenced for each sample to maximize the chances of detecting all target alleles.

5.4.13 Target sequences for functional knock outs

Wild type *NODAL* exon 1 target 1 and 2 (gRNA targets underlined):

GCCACTGCGCTCCTGCGTACGCGGGGGCAGCCCTCGTCGCCATCCCCTCTGGCGTACATGCTGAGCCTCTACCGCGACCCGCT

NODAL exon 1 target 1 knockout:

GCCAC----------AGCCCTCGTCGCCATCCCCTCTGGCGTACATGCTGAGCCTCTACCGCGACCCGCT

GCCACTGCGCTCCTG----------GGCAGCCCTCGTCGCCATCCCCTCTGGCGTACATGCTGAGCCTCTACCGCGACCCGCT

NODAL exon 1 target 2 knockout:

```
GCCACTGCGCTCCTGCGTACGCGGGGGCAGCCCTCGT-----------------------------------CGACCCGCT
GCCACTGCGCTCCTGCGTACGCGGGGGCAGCCCTCGTCGCCATCCCCTCT-GCGTACATGCTGAGCCTCTACCGCGACCCGCT
```

5.4.14 Inducible protein expression

T47D cells with stably integration of the inducible *NODAL* variant construct were treated for 96 hours with 1 µg/mL doxycycline (Sigma-Aldrich; St. Louis, Missouri, USA). For the final 24 hours, cells were cultured in the presence of serum-free RPMI media with 1 µg/mL doxycycline for collection of conditioned media. Protein was extracted from cells using mammalian protein extraction reagent (mPER; Thermo Fisher) containing the Halt Protease and Phosphatase Inhibitor Cocktail (Thermo Fisher). Lysates were incubated at room temperature for five minutes and mixed thoroughly, then centrifuged at 15,000g for 20 minutes to pellet insoluble cell debris. Protein supernatants were decanted and retained for analysis. Protein concentration was determined using the Pierce BCA Protein Assay Kit (Thermo Fisher) with a standard curve consisting of known concentrations of albumin. Corresponding conditioned media was collected and spun at 300 g for 10 minutes to eliminate floating cells and large debris. Remaining media was carefully decanted for concentration using Amicon Ultra Centrifugal Filters (Milipore) at 3,000g for 1 hour at 12°C or until media was concentrated in volume by approximately 250-fold. Halt Protease and Phosphatase Inhibitor Cocktail (Thermo Fisher) was added to concentrated conditioned media.

Samples were mixed with 4X Laemmli sample buffer (Bio-rad) containing 5% (v/v) 2-Mercaptoethanol (Sigma-Aldrich) and boiled for five minutes. SDS-PAGE was conducted with 12.5% Acrylamide gels. Precision Plus Protein Dual Color Standards (Bio-rad) were used to confirm approximate molecular weights of detected bands. Proteins were transferred to a low auto fluorescence PVDF membrane (Bio-rad) using the Trans Blot Turbo (Bio-rad) with settings of 25 V and 1.3 A for 15 minutes. After transfer, the membrane was washed briefly in PBS, and then blocked for one hour at room temperature with Odyssey Blocking Buffer (Li-Cor; Lincoln, Nebraska, USA). The membrane was incubated overnight in primary antibody solution consisting of Odyssey Blocking Buffer with 0.1% Tween-20 (Sigma-Aldrich) and mouse anti MYC-tag (9B11) antibody (#2276; Cell Signaling Technology; Massachusetts, USA) at a dilution of

1/1,000. Rabbit anti β-Tubulin polyclonal antibody (Li-Cor 926-42211) was used at a dilution of 1/1,000 as a loading control for cell lysates. Membranes were treated with corresponding Li-Cor anti mouse and anti-rabbit fluorescent secondary antibodies for one hour at room temperature at dilutions of 1/15,000 in Odyssey Blocking Buffer with 0.1% Tween-20 (Sigma-Aldrich) and 0.01% SDS (Thermo Fisher). Membranes were imaged using the Li-Cor Odyssey Clx imaging system.

5.5 References

1. University Health Network. Princess Margaret Cancer Cenre. What Is Ovarian Cancer? Retrieved October 24, 2016, from https://www.preventovariancancer.ca/what-is-ovarian-cancer
2. Quail, D. F., Siegers, G. M., Jewer, M., & Postovit, L.-M. (2013). Nodal signalling in embryogenesis and tumourigenesis. *The International Journal of Biochemistry & Cell Biology*, *45*(4), 885–898. doi:10.1016/j.biocel.2012.12.021
3. Bodenstine, T. M., Chandler, G. S., Seftor, R. E. B., Seftor, E. A., & Hendrix, M. J. C. (2016). Plasticity underlies tumor progression: role of Nodal signaling. *Cancer and Metastasis Reviews*, *35*(1), 21–39. doi:10.1007/s10555-016-9605-5
4. Fu, G., & Peng, C. (2011). Nodal enhances the activity of FoxO3a and its synergistic interaction with Smads to regulate cyclin G2 transcription in ovarian cancer cells. *Oncogene*, *30*(37), 3953–3966. doi:10.1038/onc.2011.127
5. Xu, G., Zhong, Y., Munir, S., Yang, B. B., Tsang, B. K., & Peng, C. (2004). Nodal Induces Apoptosis and Inhibits Proliferation in Human Epithelial Ovarian Cancer Cells via Activin Receptor-Like Kinase 7. *The Journal of Clinical Endocrinology & Metabolism*, *89*(11), 5523–5534. doi:10.1210/jc.2004-0893
6. Ye, G., Fu, G., Cui, S., Zhao, S., Bernaudo, S., Bai, Y., et al. (2011). MicroRNA 376c enhances ovarian cancer cell survival by targeting activin receptor-like kinase 7: implications for chemoresistance. *Journal of Cell Science*, *124*(3), 359–368. doi:10.1242/jcs.072223
7. Hardy, K. M., Strizzi, L., Margaryan, N. V., Gupta, K., Murphy, G. F., Scolyer, R. A., & Hendrix, M. J. C. (2015). Targeting Nodal in Conjunction with Dacarbazine Induces Synergistic Anticancer Effects in Metastatic Melanoma. *Molecular Cancer Research*, *13*(4), 670–680. doi:10.1158/1541-7786.MCR-14-0077
8. Lonardo, E., Hermann, P. C., Mueller, M.-T., Huber, S., Balic, A., Miranda-Lorenzo, I., et al. (2011). Nodal/Activin Signaling Drives Self-Renewal and Tumorigenicity of Pancreatic Cancer Stem Cells and Provides a Target for Combined Drug Therapy. *Cell Stem Cell*, *9*(5), 433–446.

doi:10.1016/j.stem.2011.10.001

9. Torsvik, A., Stieber, D., Enger, P. Ø., Golebiewska, A., Molven, A., Svendsen, A., et al. (2014). U-251 revisited: genetic drift and phenotypic consequences of long-term cultures of glioblastoma cells. *Cancer medicine*, *3*(4), 812–824. doi:10.1002/cam4.219

10. Masramon, L., Vendrell, E., Tarafa, G., Capellà, G., Miró, R., Ribas, M., & Peinado, M. A. (2006). Genetic instability and divergence of clonal populations in colon cancer cells in vitro. *Journal of Cell Science*, *119*(Pt 8), 1477–1482. doi:10.1242/jcs.02871

11. Kleensang, A., Vantangoli, M. M., Odwin-DaCosta, S., Andersen, M. E., Boekelheide, K., Bouhifd, M., et al. (2016). Genetic variability in a frozen batch of MCF-7 cells invisible in routine authentication affecting cell function. *Scientific Reports*, *6*, 28994. doi:10.1038/srep28994

12. van Boxtel, A. L., Chesebro, J. E., Heliot, C., Ramel, M.-C., Stone, R. K., & Hill, C. S. (2015). A Temporal Window for Signal Activation Dictates the Dimensions of a Nodal Signaling Domain. *Developmental Cell*, *35*(2), 175–185. doi:10.1016/j.devcel.2015.09.014

13. Robertson, E. J. (2014). Dose-dependent Nodal/Smad signals pattern the early mouse embryo. *Seminars in Cell & Developmental Biology*, *32*, 73–79. doi:10.1016/j.semcdb.2014.03.028

14. Chen, C., & Shen, M. M. (2004). Two Modes by which Lefty Proteins Inhibit Nodal Signaling. *Current Biology*, *14*(7), 618–624. doi:10.1016/j.cub.2004.02.042

15. Müller, P., Rogers, K. W., Jordan, B. M., Lee, J. S., Robson, D., Ramanathan, S., & Schier, A. F. (2012). Differential diffusivity of Nodal and Lefty underlies a reaction-diffusion patterning system. *Science*, *336*(6082), 721–724. doi:10.1126/science.1221920

16. Sakuma, R., Ohnishi Yi, Y.-I., Meno, C., Fujii, H., Juan, H., Takeuchi, J., et al. (2002). Inhibition of Nodal signalling by Lefty mediated through interaction with common receptors and efficient diffusion. *Genes to cells : devoted to molecular & cellular mechanisms*, *7*(4), 401–412.

17. Juan, H., & Hamada, H. (2001). Roles of nodal-lefty regulatory loops in embryonic patterning of vertebrates. *Genes to cells : devoted to molecular & cellular mechanisms*, *6*(11), 923–930.

18. Papageorgiou, I., Nicholls, P. K., Wang, F., Lackmann, M., Makanji, Y., Salamonsen, L. A., et al. (2009). Expression of nodal signalling components in cycling human endometrium and in endometrial cancer. *Reproductive Biology and Endocrinology*, *7*(1), 122–11. doi:10.1186/1477-7827-7-122

19. Postovit, L.-M., Margaryan, N. V., Seftor, E. A., Kirschmann, D. A., Lipavsky, A., Wheaton, W. W., et al. (2008). Human embryonic stem cell microenvironment suppresses the tumorigenic phenotype of aggressive cancer cells. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(11), 4329–4334. doi:10.1073/pnas.0800467105

20. Joung, J. K., & Sander, J. D. (2012). TALENs: a widely applicable technology for targeted genome editing. *Nature reviews. Molecular cell biology*, *14*(1), 49–55. doi:10.1038/nrm3486

21. Sun, N., & Zhao, H. (2013). Transcription activator-like effector nucleases (TALENs): A highly efficient and versatile tool for genome editing. *Biotechnology and Bioengineering*, n/a–n/a. doi:10.1002/bit.24890

22. Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, *8*(11), 2281–2308. doi:10.1038/nprot.2013.143

23. Doudna, J. A., & Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science*, *346*(6213), 1258096–1258096. doi:10.1126/science.1258096

24. Kim, Y., Kweon, J., & Kim, J.-S. (2013). TALENs and ZFNs are associated with different mutation signatures. *Nature Methods*, *10*(3), 185. doi:10.1038/nmeth.2364

25. Yi, L., & Li, J. (2016). CRISPR-Cas9 therapeutics in cancer: promising strategies and present challenges. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, *1866*(2), 197–207. doi:10.1016/j.bbcan.2016.09.002

26. Reardon, S. (2016). First CRISPR clinical trial gets green light from US panel. *Nature*. doi:10.1038/nature.2016.20137

27. Yang, Z., Steentoft, C., Hauge, C., Hansen, L., Thomsen, A. L., Niola, F., et al. (2015). Fast and sensitive detection of indels induced by precise gene targeting. *Nucleic Acids Research*, *43*(9), e59–e59. doi:10.1093/nar/gkv126

28. Yu, C., Zhang, Y., Yao, S., & Wei, Y. (2014). A PCR Based Protocol for Detecting Indel Mutations Induced by TALENs and CRISPR/Cas9 in Zebrafish. *PLoS ONE*, *9*(6), e98282–7. doi:10.1371/journal.pone.0098282

29. Hendel, A., Fine, E. J., Bao, G., & Porteus, M. H. (2015). Quantifying on- and off-target genome editing. *Trends in biotechnology*, *33*(2), 132–140. doi:10.1016/j.tibtech.2014.12.001

30. Wang, K., Mei, D. Y., Liu, Q. N., Qiao, X. H., Ruan, W. M., Huang, T., & Cao, G. S. (2015). Research of methods to detect genomic mutations induced by

CRISPR/Cas systems. *Journal of biotechnology*, *214*, 128–132. doi:10.1016/j.jbiotec.2015.09.029

31. Vouillot, L., Thélie, A., & Pollet, N. (2015). Comparison of T7E1 and surveyor mismatch cleavage assays to detect mutations triggered by engineered nucleases. *G3 (Bethesda, Md.)*, *5*(3), 407–415. doi:10.1534/g3.114.015834

32. Mock, U., Machowicz, R., Hauber, I., Horn, S., Abramowski, P., Berdien, B., et al. (2015). mRNA transfection of a novel TAL effector nuclease (TALEN) facilitates efficient knockout of HIV co-receptor CCR5. *Nucleic Acids Research*, *43*(11), 5560–5571. doi:10.1093/nar/gkv469

33. Miyaoka, Y., Chan, A. H., Judge, L. M., Yoo, J., Huang, M., Nguyen, T. D., et al. (2014). Isolation of single-base genome-edited human iPS cells without antibiotic selection. *Nature Methods*, *11*(3), 291–293. doi:10.1038/nmeth.2840

34. Milot, E., & Ellis, J. (2006). Transgene Silencing. In *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine* (pp. 1896–1899). Springer Berlin Heidelberg. doi:10.1007/3-540-29623-9_2940

35. Qian, K., Huang, C.-L., Chen, H., Blackbourn, L. W., IV, Chen, Y., Cao, J., et al. (2014). A Simple and Efficient System for Regulating Gene Expression in Human Pluripotent Stem Cells and Derivatives. *STEM CELLS*, *32*(5), 1230–1238. doi:10.1002/stem.1653

36. Doyle, E. L., Booher, N. J., Standage, D. S., Voytas, D. F., Brendel, V. P., VanDyk, J. K., & Bogdanove, A. J. (2012). TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. *Nucleic Acids Research*, *40*(W1), W117–W122. doi:10.1093/nar/gks608

37. Streubel, J., Blücher, C., Landgraf, A., & Boch, J. (2012). TAL effector RVD specificities and efficiencies. *Nature Biotechnology*, *30*(7), 593–595. doi:10.1038/nbt.2304

38. Cong, L., Zhou, R., Kuo, Y.-C., Cunniff, M., & Zhang, F. (2012). Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nature Communications*, *3*, 968. doi:10.1038/ncomms1962

39. Cermak, T., Doyle, E. L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., et al. (2011). Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Research*, *39*(12), e82–e82. doi:10.1093/nar/gkr218

40. Jankele, R., & Svoboda, P. (2014). TAL effectors: tools for DNA Targeting. *Briefings in Functional Genomics*, *13*(5), 409–419. doi:10.1093/bfgp/elu013

41. Welch, D. R., Bisi, J. E., Miller, B. E., Conaway, D., Seftor, E. A., Yohem, K. H.,

et al. (1991). Characterization of a highly invasive and spontaneously metastatic human malignant melanoma cell line. *International Journal of Cancer*, *47*(2), 227–237.

42. Franken, N. A. P., Rodermond, H. M., Stap, J., Haveman, J., & van Bree, C. (2006). Clonogenic assay of cells in vitro. *Nature Protocols*, *1*(5), 2315–2319. doi:10.1038/nprot.2006.339

43. Frank, S., Skryabin, B. V., & Greber, B. (2013). A modified TALEN-based system for robust generation of knock-out human pluripotent stem cell lines and disease models. *BMC Genomics*, *14*(1), 773. doi:10.1186/1471-2164-14-773

44. Lin, Y., Cradick, T. J., & Bao, G. (2014). Designing and Testing the Activities of TAL Effector Nucleases. In *Gene Correction* (Vol. 1114, pp. 203–219). Totowa, NJ: Humana Press. doi:10.1007/978-1-62703-761-7_13

45. Corbisier, P., Pinheiro, L., Mazoua, S., Kortekaas, A.-M., Chung, P. Y. J., Gerganova, T., et al. (2015). DNA copy number concentration measured by digital and droplet digital quantitative PCR using certified reference materials. *Analytical and Bioanalytical Chemistry*, *407*(7), 1831–1840. doi:10.1007/s00216-015-8458-z

Chapter 6

6 Overall discussion

6.1 Complexity of human gene expression

The work presented in this thesis suggests that there is a great deal of molecular complexity for human *NODAL* gene expression at multiple levels. Specifically, I discovered expression of multiple transcripts transcribed from the *NODAL* locus, confirming both alternative transcriptional start site usage and alternative splicing of *NODAL* transcripts. Alternative splicing of *NODAL* is genetically regulated, and the translated protein product of the novel full-length transcript is subject to differential N-glycosylation and is functionally distinct from constitutively spliced *NODAL*. This complexity was previously unknown and can now be incorporated into and enrich experimental models of human *NODAL* function. These details will also help refine the development and evaluation of inhibitors of *NODAL* signalling desirable for potential targeted therapy in cancer.

6.2 Discovery and characterization of a human-specific alternatively spliced *NODAL* transcript

At the genomic level, I identified a functional non-coding polymorphism in *NODAL*'s second intron. This SNP directly controls the novel alternative splicing of *NODAL* transcripts, resulting in expression of the first identified *NODAL* transcript variant. Thus, *NODAL* is differentially alternatively spliced between individuals. I also extensively characterized this transcript variant and the constitutive *NODAL* isoform at the RNA and protein levels. The alternative *NODAL* exon contributed to a full-length *NODAL* variant transcript containing a slightly truncated open reading frame (ORF) relative to constitutive *NODAL*. Most *NODAL* variant transcripts contained transcriptional start sites corresponding to constitutive exon 1 as was the case for total *NODAL*. However, a minority of *NODAL* variant transcripts did not contain constitutive exon 1, but instead spliced directly from constitutive exon 2 to a novel upstream first exon. I found evidence of this exon's preferential splicing in *NODAL* variant transcripts relative to total *NODAL*.

This transcript was very rare, but is indicative of potential further transcript complexity for human *NODAL* beyond what was examined here. The 3' ends of *NODAL* variant transcripts were defined by polyadenylation guided by the same two PAS found for total *NODAL* transcripts. However, the *NODAL* variant transcripts were much more likely to utilize the more distal PAS than *NODAL* transcripts in general.

While the assays used provided confidence in the nature of 3' transcript ends, exact determination of 5' ends was difficult. This is a limitation of reverse transcription-based methods of 5' end determination such as RACE, as incomplete reverse transcription can result in a collection of 5' truncated cDNAs that are indistinguishable from true 5' termini. This was evident for analysis of *NODAL* variant transcripts. Future studies will use the RLM-RACE technique that specifically adds an adapter oligo to capped 5' mRNA ends providing the specificity required for true 5' end determination [1]. Collectively, these differences are indicative of coordinated regulation of *NODAL* variant processing, beyond the genetic modulation of a splice donor site in cis.

Our work illustrated how oft-overlooked genetic polymorphisms can play important roles in gene expression, specifically splicing. It also distinguished true alternative splicing from allele-specific expression, developed improved assays for precise quantification of alternatively spliced transcripts, and investigated the full-length nature of such transcripts. These are all aspects typically lacking in conventional analyses of alternative splicing events. Specifically, the interrogation of full-length transcripts containing open reading frames will prove important in identifying potential proteoforms translated from transcripts subject to AS. On a genome-wide scale, the investigation of alternatively spliced ORF-containing transcripts has been extremely limited. Only very recently, a study attempted to begin to characterize alternatively spliced transcripts contributing to what they term the "ORFeome"—the full collection of open reading frame-containing transcripts expressed by the human genome [2]. Using a targeted approach to clone and sequence a subset of human transcripts of interest from five pooled tissue samples, they found a total of 917 alternatively spliced transcripts from 506 corresponding reference transcripts for 506 genes. Notably, while only 11% of the exon-exon junctions contained within these transcripts were novel, a staggering 70% of the full-length isoforms found

had never been curated in any of the databases examined and were thus completely novel. This served as a striking indication of how little is known about the potential for AS to promote proteomic diversity, even at the RNA level.

Differences in the regulation of *NODAL* splice variant gene expression at the level of protein were also apparent. First, the NODAL proteoforms resulting from translation of the alternatively spliced open reading frames were differentially secreted into the media. This capacity was enhanced for the NODAL variant relative to constitutive NODAL. A differential banding pattern between the processed mature peptides of the two proteoforms in the conditioned media was found to result from differential N-glycosylation. Interestingly, the novel C-terminal N-glycosylation of the NODAL variant was similar in nature to some Nodals and other TGF-beta superfamily ligands in non-human organisms [3]. Le Good and colleagues found that artificial introduction of an analogous N-glycosylation motif into the mature domain of constitutive Nodal resulted in increased Nodal stability and corresponding signalling range. The work presented in this thesis suggests that N-glycosylation of the NODAL variant mature peptide does not have a general stabilizing effect, though it does promote increased secretion relative to constitutive NODAL.

6.3 *NODAL* expression in human cancer cell lines and embryonic stem cells

It was unclear if *NODAL* is similarly alternatively spliced in cancer, owing to the surprising discovery that *NODAL* transcripts were detectable at extremely low levels in cancer cell lines. This is in apparent contradiction with numerous functional studies employing *NODAL* knockdown suggesting that *NODAL* is expressed. Careful quantitative parallel analysis of *NODAL* transcript and corresponding NODAL protein levels from the same cultures will be required to determine if there is a consistent discrepancy between transcript and protein levels. Alternatively, it is possible that the low levels detected here result from genuine low gene expression and reflect a tendency for *NODAL* expression to drift between isolates of the same cell lines. If the former is true, several mechanisms could be responsible. It is possible that *NODAL* transcripts are translated very efficiently, allowing substantial protein to accumulate from a limited

number of transcripts. Large pools of *NODAL* pre-mRNA may also be rapidly spliced, translated, and broken down, accounting for a very transient fully-spliced transcript pool. It is also possible that *NODAL* transcripts contain different 5' and 3' untranslated regions relative to transcripts detected in human embryonic stem cells, and that these untranslated regions confer complex structure that does not permit efficient reverse transcription.

The work presented here is the first to directly compare multiple *NODAL* primer probe assays across multiple samples using absolutely quantitative ddPCR. Consequently, this work demonstrated that a natural antisense transcript transcribed from the human *NODAL* locus contains all of *NODAL*'s constitutive exon 2 sequence. This implies that transcripts detected with constitutive exon 2 assays are not specific for *NODAL*: amplicons derived from this region could result from amplification of both *NODAL* and the antisense transcript. Moreover, the resulting amplicons would be indistinguishable in sequence analyses as they would both align perfectly with the intended *NODAL* target. Going forward, assays either spanning an exon-exon boundary, or outside of constitutive exon 2 should be used to assess *NODAL* transcript levels. It will also be of interest to specifically interrogate pre-mRNA with assays that amplify across intron-exon boundaries to determine if abundant unspliced *NODAL* transcript may also explain discrepancies between junction spanning assays and those exclusively within constitutive exon 2. The accumulation of *Nodal* pre-mRNA has been described in zebrafish [4].

In addition to low transcript levels in cancer systems, low transcript levels were also variably found in human embryonic stem cells. This was not primarily the result of technical inefficiencies, and existed despite cells displaying classic pluripotent stem cell morphology, and expression of pluripotency markers. *NODAL* expression has been shown to decrease very quickly upon spontaneous differentiation of hES cells in suboptimal culture conditions [5]. Therefore, it is possible that low *NODAL* levels were an indication of early differentiation taking place, not yet apparent at the level of cell morphology. However, the magnitude of the difference in *NODAL* levels between "high" and "low" samples was much greater than the decrease in *NODAL* levels reported in [5] after prolonged differentiation. Another possibility is that pluripotent stem cells can be propagated in distinct pluripotent states, and that the microenvironment influences which

state is preferred. This work demonstrated that choice of culture media had a dramatic effect on *NODAL* gene expression. Both MEF-conditioned media and defined media have been well established in their abilities to support hES cell self-renewal and pluripotency [6]. Cell transfer to MEF-CM dramatically increased *NODAL* transcript levels. This effect was also reversible. Perhaps surprisingly, *NODAL* levels were lower in defined mTESR-1 media that is less prone to the batch variation and biological heterogeneity that accompany the use of secreted factors from MEFs. Since defined media such as mTESR-1 contains TGF-beta, it is possible that this supplementation satisfies the cell's requirement for active NODAL/Activin/TGF-beta signalling through SMAD2/3, and that endogenous expression of *NODAL* is not strictly required and thus subject to drift.

6.4 NODAL variant function

The *NODAL* variant ORF lacked the canonical activity of constitutive *NODAL* in an endogenous zebrafish embryo reporter system. However, over expression of the *NODAL* variant induced stem cell-like phenotypes in ovarian cancer in a similar fashion to constitutive *NODAL*, although to a lesser extent. This indicated that the *NODAL* variant may be functional in a less well-regulated cancerous context. It is unclear whether this apparent function of the *NODAL* variant was related to *NODAL* variant-specific activity, or resulted from activity of the peptide sequence shared with constitutive NODAL. Truncation mutants lacking the novel C-terminal NODAL variant region have been constructed that will be used to distinguish between these two possibilities. One possibility is that the NODAL pro-domain has function independent of the mature peptide after proteolytic cleavage, although this has not yet been reported. Future work will examine if the NODAL pro-peptide common to both proteoforms has independent functions, such as stabilizing endogenous NODAL or binding with GDF1 or other TGF-beta ligands. These possible functions were not sufficient to induce a bona fide canonical NODAL signalling response in the zebrafish embryo model used. However, it should be noted that this model focused on signaling events and development only up until the early stages of gastrulation. It is possible that over-expression of the NODAL pro-peptide can impact gene expression and development at later stages of development, such as during the establishment of left-right asymmetry and subsequent organ development. Beyond

development, the pro-domain may also be functionally relevant in cancer systems. Future work will also examine if *NODAL* and *NODAL* variant effects on cancer cell phenotypes are strictly dependent on EGF-CFC co-receptors such as Cripto.

This work also demonstrated possible coordinated regulation of the two *NODAL* isoforms in hES cells. *NODAL* variant specific knockdown resulted in similar changes in gene expression induced upon total *NODAL* knockdown. It is possible this is indicative of partially redundant function between the two isoforms. However, the sequence and functional divergence of the two splice variants in the zebrafish signalling assay suggest this is unlikely to be broadly true. This work demonstrated that knockdown of the *NODAL* variant resulted in a proportional decrease in constitutive *NODAL* expression. It is possible that this effect is direct and that interfering with *NODAL* variant splicing had a general effect on the processing of all *NODAL* transcripts. For example, it is possible that the alternative exon splice donor site locus plays a role in constitutive *NODAL* splicing (perhaps as an intronic splicing enhancer) independent of SNP rs2231947 genotype. Future work will test this hypothesis by treating a homozygous C|C cell line such as H1 with the morpholino targeting the alternative exon splice site to see if a similar effect on constitutive *NODAL* transcript levels is observed.

In general, this work did not find any negative impact of *NODAL* variant expression. Although the allele from which the *NODAL* variant was spliced in a rs2231947-heterozygous hES cell line was responsible for slightly less production of processed *NODAL* transcript in general, *NODAL* variant splicing did not preclude productive processing of constitutive *NODAL*: *NODAL* variant-specific knockdown in fact reduced constitutive *NODAL* levels, indicative of a putative supporting effect on *NODAL* expression. Furthermore, even when co-injected in excess, the *NODAL* variant open reading frame did not have a dominant negative effect on the robust signalling response induced by constitutive *NODAL* in zebrafish embryos. Indeed, any potential deleterious effect of *NODAL* variant expression is likely limited in scope as high rates of the T allele for rs2231947 are found in numerous populations of adults that presumably experienced healthy development.

However, the genetic associations for SNP rs2231947 on characteristics of hES cell lines were striking. This suggests that any harmful impact of the linkage group marked by rs2231947 is likely compensated for in normal development, but is manifested in *ex vivo* hES cell models that have been displaced from their endogenous microenvironment. If *NODAL* variant splicing is responsible for selection against prospective male human embryonic stem cell lines and the X inactivation process in female hES cell lines, these effects are likely achieved through mechanisms and contexts that are not easily recapitulated experimentally. Robust and inducible *NODAL* variant over-expression in hES cells will be used to determine how this *NODAL* transcript potentially impacts pluripotency. In addition, we identified another putative functional polymorphism in an upstream *NODAL* enhancer element in high LD with rs2231947 displaying the same associations. Future work to endogenously manipulate combinations of alleles for this SNP as well as rs2231947 will examine potential combinatorial effects on *NODAL* gene expression. These *NODAL* SNPs and *NODAL* in general can also be incorporated into models of X chromosome inactivation in human embryonic stem cells.

This work did not directly assess endogenous translation of the *NODAL* variant. Custom antibodies were generated against the unique C-terminal region of the NODAL variant proteoform. However, these antibodies were generally non-specific and it was difficult to consistently obtain samples with robust *NODAL* expression.

6.5 Novel aspects of constitutive *NODAL* biology

Beyond characterization of the novel *NODAL* splice variant, I also detailed many aspects of constitutive *NODAL* transcript and protein. *NODAL* was found to be alternatively polyadenylated in hES cells. If subcultures of cancer cell lines expressing high levels of *NODAL* could be obtained, it would be interesting to see if a skewed pattern of *NODAL* alternative polyadenylation exists in a cancer context. It would also be interesting to explore if alternative polyadenylation influences translation efficiency or miRNA targeting, as direct regulation of *NODAL* by endogenous RNA interference has not been described in humans.

This work also explored a novel relationship between NODAL processing and dimerization. Conservative mutation of the putative interchain disulfide bond-forming C312 residue dramatically affected the processing of secreted NODAL, increasing the mature:full-length peptide ratio. This was true for C-terminal tagged NODAL proteins despite lack of abundant putative homo-dimerization in the conditioned media. Future work will assess to what degree C312 mutation disrupts NODAL homo-dimerization, and whether proteolytic enzymes such as PACE4 preferentially cleave monomer NODAL ligands relative to their dimeric counterparts. The exact nature of the interchain disulfide bond complexes can also be confirmed using mass spectrometry-based techniques [7].

6.6 Novel transcripts originating from the *NODAL* gene locus

Beyond characterization of the full-length *NODAL* isoforms, this work also identified two other novel transcripts originating from the *NODAL* locus. Namely, these consisted of a natural antisense transcript encompassing constitutive exon 2, and a circular exon formed by a back-splice of the constitutive exon 2 splice donor to its own splice acceptor site. Future work will assess the function of these novel transcripts. For example, does their over-expression affect linearly spliced full-length *NODAL* transcripts? And how do these novel transcripts respond to microenvironmental changes such as hypoxia relative to full-length *NODAL*?

Indeed, it is also still possible that other *NODAL* isoforms and proteoforms exist beyond those detailed here. *NODAL* expression may differ in individuals, cell types, stages of development, and microenvironments not studied here. Furthermore, many of the assays used here are biased by what is already known about *NODAL* gene expression and may not have been sensitive to detection of currently unidentified *NODAL* molecules. For example, RACE analyses that employed primers targeting the second constitutive exon of *NODAL* could not detect potential transcripts where exon 2 is skipped. The targeted and gene specific methodologies used here provided excellent sensitivity that may exceed genome-wide methods where extensive filtering of data may be required for efficient and confident analyses. However, both types of studies can complement each other to provide a rich view of gene expression from a locus of interest. For example, whole transcriptome

shotgun sequencing would be able to identify the potential *NODAL* transcripts lacking constitutive exon 2.

6.7 Widespread complexity in gene expression

Although this work was focused on one gene, it is likely that the complexity uncovered at the *NODAL* gene locus is typical of many protein coding genes, rather than an exceptional case. This is supported by a tremendous amount of progress that has been made in recent decades in detailing the complexity of gene expression on a genome-wide scale [8]. At the RNA level, this complexity is generated by alternative transcriptional start sites, alternative splicing, and alternative polyadenylation, all of which were demonstrated here for human *NODAL*, and are each now estimated to affect gene expression for the majority of protein coding genes [9]. However, on an individual gene basis, how this alternative processing affects gene expression and function often remains completely unknown. This is perhaps the result of difficulty in detecting some alternatively processed transcripts, or difficulty incorporating them into conventional models used to assess gene function. Going forward, extensive characterization of these transcripts will enrich our understanding of countless genes, as numerous examples of alternatively processed transcripts already have. To what extent alternatively processed transcripts are translated into functional proteins remains unclear [10] and has been the source of some controversy [11]. Skeptics suggest that alternative processing of transcripts may largely represent biological noise, as the majority of alternative processing events are not well-conserved across species. The alternate argument is that extensive alternative RNA processing is one mechanism to generate proteomic diversity and mediate many important inter-species differences in development and physiology.

6.8 Combinatorial complexity of gene expression

Collectively, the effect of numerous points of regulation of gene expression on diversity at the protein level is quite staggering. It has been proposed that distinct protein products of the same gene locus resulting from allelic variation within or between individuals, as well as processes such as alternative splicing and post-translational modifications, be termed “proteoforms” [12]. This term is analogous to the term “isoform” for nucleic

acids. Collectively, the full array of proteoforms expressed from loci throughout the genome contribute to the proteome. One general process responsible for a great deal of proteomic diversity is post-translational modification (PTM). Over 100 distinct types of PTMs have been characterized (reviewed in [13]), and these modifications can be differentially applied to multiple amino acids of a single peptide. Collectively, it has been estimated that processes generating proteomic diversity account for the generation of an estimated 1 million proteoforms from a genome of only approximately 20,000 protein-coding genes [14]. This implies an average of roughly 50 proteoforms from every gene locus. This number is given perspective once the combinatorial effects of different regulatory processes are considered. The diversity of *NODAL* gene products reported here will be used to illustrate: *NODAL* is alternatively spliced to yield two distinct proteins. Each of these proteins was subject to alternative N-glycosylation in the pro-domain. Even alone, these two simple points of regulation account for 2 x 4 = 8 potential proteoforms if each combination of unmodified, N72 N-glyc only, N199 N-glyc only, and N72 & N199 N-glyc are considered. If an individual heterozygous for a common SNP that results in a single amino acid change such as rs1904589 is considered, this number quickly doubles to 16. All this diversity has important implications for protein function and can change dynamically with context and between individuals. Specific detection and characterization of distinct proteoforms is a major challenge currently facing research in molecular biology.

6.9 Conclusion

Perhaps the most profound and widely applicable finding of this work is that a single nucleotide polymorphism can have a substantial yet relatively benign impact on gene expression and function. The study of genetic polymorphisms or mutations has traditionally been approached from a pathogenic perspective. The mindset is that mutations often face substantial negative selection pressure, and generally disrupt "normal" expression and function of proteins, possibly conferring a disease. While this is undoubtedly true of some variants, the impact of the vast majority of genetic polymorphisms, especially those that are common, remains almost completely unknown. The *NODAL* SNP rs2231947 studied here illustrates how a common polymorphism can

affect both the expression and function of a highly conserved protein-coding gene with essential roles in early embryonic development. The high frequency of this polymorphic allele in multiple human populations suggests that its impact on *NODAL* gene expression is likely well tolerated endogenously, and not deleterious to the development of those individuals carrying it.

The impact of this non-coding SNP is one example of how genetic heterogeneity between individuals impacts the molecular biology of the cell. It is unlikely that *NODAL* is an exceptional gene in this sense. Thus, the overall approach of this thesis can serve as a framework for the study of complexity at multiple levels for any protein-coding gene of interest. While there are currently not many functional annotations for non-coding polymorphisms, as we continue to move away from a protein-coding gene-centric view of molecular biology, such annotations will undoubtedly be major contributors to functional annotation of the human genome. This will allow for a more nuanced view of molecular biology in general, as we strive to understand not only differences between individuals, but also what makes us human.

6.10 References

1. Chen, N., Wang, W.-M., & Wang, H.-L. (2016). An efficient full-length cDNA amplification strategy based on bioinformatics technology and multiplexed PCR methods. *Scientific Reports*, *5*, 19420. doi:10.1038/srep19420
2. Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., et al. (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*, *164*(4), 805–817. doi:10.1016/j.cell.2016.01.029
3. Le Good, J. A., Joubin, K., Giraldez, A. J., Ben-Haim, N., Beck, S., Chen, Y., et al. (2005). Nodal Stability Determines Signaling Range. *Current Biology*, *15*(1), 31–36. doi:10.1016/j.cub.2004.12.062
4. Sampath, K., & Robertson, E. J. (2016). Keeping a lid on nodal: transcriptional and translational repression of nodal signalling. *Open Biology*, *6*(1), 150200–8. doi:10.1098/rsob.150200
5. Besser, D. (2004). Expression of Nodal, Lefty-A, and Lefty-B in Undifferentiated Human Embryonic Stem Cells Requires Activation of Smad2/3. *Journal of Biological Chemistry*, *279*(43), 45076–45084. doi:10.1074/jbc.M404979200

6. Desai, N., Rambhia, P., & Gishto, A. (2015). Human embryonic stem cell cultivation: historical perspective and evolution of xeno-free culture systems. *Reproductive Biology and Endocrinology*, *13*(1), 9–15. doi:10.1186/s12958-015-0005-4

7. Borges, C. R., & Sherma, N. D. (2014). Techniques for the Analysis of Cysteine Sulfhydryls and Oxidative Protein Folding. *Antioxidants & Redox Signaling*, *21*(3), 511–531. doi:10.1089/ars.2013.5559

8. Niklas, K. J., Bondos, S. E., Dunker, A. K., & Newman, S. A. (2015). Rethinking gene regulatory networks in light of alternative splicing, intrinsically disordered protein domains, and post-translational modifications. *Frontiers in cell and developmental biology*, *3*(37), 8. doi:10.3389/fcell.2015.00008

9. de Klerk, E., & t Hoen, P. A. C. (2015). Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends in Genetics*, *31*(3), 128–139. doi:10.1016/j.tig.2015.01.001

10. Hegyi, H., Kalmar, L., Horvath, T., & Tompa, P. (2011). Verification of alternative splicing variants based on domain integrity, truncation length and intrinsic protein disorder. *Nucleic Acids Research*, *39*(4), 1208–1219. doi:10.1093/nar/gkq843

11. Tress, M. L., Abascal, F., & Valencia, A. (2016). Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends in Biochemical Sciences*, 1–13. doi:10.1016/j.tibs.2016.08.008

12. Smith, L. M., Kelleher, N. L., Linial, M., Goodlett, D., Langridge-Smith, P., Ah Goo, Y., et al. (2013). Proteoform: a single term describing protein complexity. *Nature Methods*, *10*(3), 186–187. doi:10.1038/nmeth.2369

13. Lichti, C. F., Wildburger, N. C., Emmett, M. R., Mostovenko, E., Shavkunov, A. S., Strain, S. K., & Nilsson, C. L. (2014). Post-translational Modifications in the Human Proteome. In *Genomics and Proteomics for Clinical Discovery and Development* (Vol. 6, pp. 101–136). Dordrecht: Springer Netherlands. doi:10.1007/978-94-017-9202-8_6

14. Nørregaard Jensen, O. (2004). Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Current Opinion in Chemical Biology*, *8*(1), 33–41. doi:10.1016/j.cbpa.2003.12.009

Appendices

Appendix A: Annotations and sequences

For what I refer to as "constitutive NODAL":

UniProtKB/Swiss-Prot ID:	Q96S42 (NODAL_HUMAN)
Ensemble/ Genocde transcript ID:	ENST00000287139/ ENST00000287139.6
NCBI Refseq:	NM_018055.4
NCBI Protein:	NP_060525.3
UCSC ID (hg38):	uc001jrc.3
UCSC ID (hg19):	uc001jrc.2

Another NODAL transcript is also present in several databases. During writing of this thesis, this transcript was curated into the RefSeq and UCSC genome browser databases. This NODAL transcript has an alternative first exon upstream of annotated NODAL exon 1, and utilizes the same exon 2 and exon 3 as annotated NODAL. This transcript was not directly assessed in this thesis, but it was not detected in 5' RACE analysis.

UCSC Genome Browser ID (hg38†):	uc057tvn.1
Ensemble/ Gencode transcript ID:	ENST00000414871/ ENST00000414871.1
NCBI RefSeq:	NM_001329906.1
UniProtKB/TrEMBL* ID:	H7C0E4 (H7C0E4_HUMAN)

† note that there is no alternative first exon NODAL record in hg19.

* note that this version of the UniProt database contains entries that have not been manually reviewed for inclusion in the UniProtKB/Swiss-Prot database.

The NODAL variant has not been annotated into any databases. The following is the sequence and genomic coordinates for the NODAL cassette alternative exon (sense strand):

>Hg38_chr10:70434100-70434215
gtggccctgccctgctgtccaaggtcatatgggaccaaaatgttttcattttactccatgaagtctggaatgagaatttcttggacttg
caatatatcctccatgccaagcctcag

For what I refer to as the NODAL natural antisense transcript:

UCSC Genome Browser ID (Hg38): RP11-104F15.9

Ensemble/ Gencode transcript ID: ENST00000624563.1

NCBI GenBank (example): AK001176.1

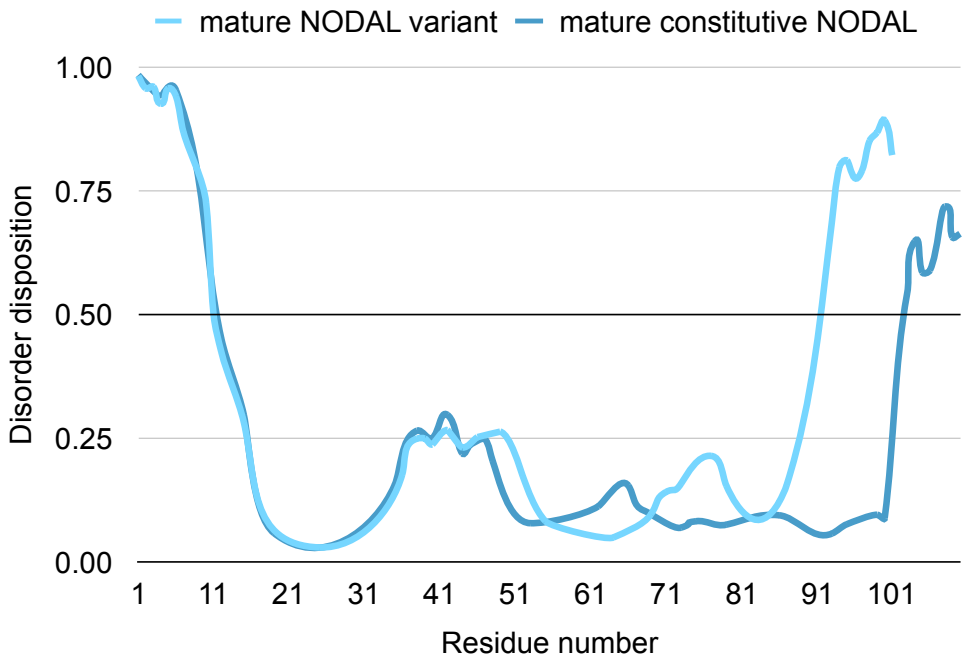
Predicted corresponding protein product (predicted signal peptide underlined):

>BAA91534.1_unnamed_protein_product_[Homo sapiens]

MVGRMKLLPNRIRTLALTAIGVVLLGVDDPGAPSDQVEVHLELDLPTQLTSVWQ
VMSTVPLAPLPGQLSLLGPPGALGFPQQGGPTQLPLLLREVGVEHKEHIGGRRCG
GPRPALSSYPGHLLLQGPRALQPLGERPGHLQNHAAQGKGDLGQSDSE

Appendix B: Additional predictions for NODAL proteins.

Appendix B1: Predicted intrinsic disorder of constitutive NODAL and NODAL variant mature peptides.



Appendix B2: Prosite PKC phosphorylation sites specific to the NODAL variant mature peptide not present in constitutive NODAL:

Pattern-ID: PKC_PHOSPHO_SITE PS00005 PDOC00005

Pattern-DE: Protein kinase C phosphorylation site

Pattern: [ST].[RK]

Sites: 79 SMK

97 SLR

Appendix C: Guidelines for precision genome editing

Appendix C1: Instructions for one-step gRNA cloning into all-in-one CRISPR/Cas9 LacZ:

1) Design an appropriate gRNA for your target of interest using software of choice.

2) Order a g-Block (IDT DNA) that includes your gRNA sequence. Replace all of the 20 "N" bases an 18-20 bp gRNA sequence in the 5'-3' orientation (as supplied by the design tools). Include a 5' G if desired. Do not leave any "N" bases in the sequence. Double check that this new sequence does not introduce a new BmsBI restriction site ("CGTCTC" or "GAGACG"). This is unlikely, but would interfere with cloning. There should be only two such sites in the whole g-Block sequence.

>example_g_block_insert_for_crispr_all_in_one

ATATATCGTCTCGAACTTGAAAGTATTTCGATTTCTTGGGTTTATATATCTTGT
GGAAAGGACGAAACACCNNNNNNNNNNNNNNNNNNNNGTTTTAGAGCTAGA
AATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACC
GAGTCGGTGCTTTTTTCTAGACACAATTGCATGAAGAATCTGCTTAGGGTTAG
GCGTTTTGCGCTAGAGACGAATTAT

3) Re-suspend g-Block in TE buffer to a final concentration of 10 ng/µL.

4) Mix the following components in a 0.2 mL PCR tube:

1)	g-Block	25 ng
2)	All-in-one CRISPR/Cas9 LacZ	75 ng
3)	BsmBI (10 U/µL)	1 µL
4)	T4 ligase (Thermo Fisher: 15224017)	1 µL
5)	T4 buffer	2 µL
6)	Nuclease-free water	to 20 µL total

5) Incubate in a standard thermal-cycler using the following conditions:

1)	37°	5 min
2)	16°	10 min

3) 37° 15 min
4) 80° 5 min

The reaction is now ready for transformation (use a maximum of 5 µL for 50 µL competent cells) and plasmid preparation. Colonies containing successfully cloned plasmids will be white if using blue/white screening. Selected clones can be sequenced using the SP6F primer.

Appendix C2: Droplet digital PCR mutation screening assay design guidelines.

ddPCR assays can be designed using Primer3Plus (http://primer3plus.com) with modified settings: 50 mM monovalent cations, 3.0 mM divalent cations, 0 mM dNTPs, and SantaLucia 1998 thermodynamic and salt correction parameters. Predicted nuclease cut sites should be positioned mid-amplicon, with 75-125 bp flanking either side up to the primer binding sites. Reference probe and primers should be designed distant from the cut site (origin of NHEJ generation). Optimal annealing temperatures should be determined empirically by temperature gradient. In general, it is recommended to design primers with T_m = 55°C, reference probes with T_m = 60°C, and NHEJ/drop-off probes with T_m = 56-57°C. However, higher melting temperatures are appropriate for high-GC targets to design primers and probes of sufficient length.

Appendix D: Copyright information

For re-use of text from:

Findlay, S.D.*, Jewer, M.*, & Postovit, L.-M. (2012). Post-transcriptional regulation in cancer progression. *Journal of Cell Communication and Signaling*, *6*(4), 233–248. doi:10.1007/s12079-012-0179-x

An application to re-print excerpts of this article in a thesis was submitted. The publisher for this copyrighted material is Springer. Permission was granted for the requested reuse under conditions specified by limited license agreement.

For re-use of text and figures from:

Findlay, S.D., & Postovit, L.-M. (2016). Common Genetic Variation in Chromosome 10 q22.1 Shows a Strong Sex Bias in Human Embryonic Stem Cell Lines and Directly Controls the Novel Alternative Splicing of Human NODAL which is Associated with XIST Expression in Female Cell Lines. *STEM CELLS*, 34(3), 791–796. doi:10.1002/stem.2258

An application to re-print this article in a thesis was submitted with the following confirmation:

"You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, and any CONTENT (PDF or image file) purchased as part of your order, is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission."

For re-use of text and figures from:

Findlay, S.D.*, Vincent, K. M.*, Berman, J. R., & Postovit, L.-M. (2016). A Digital PCR-Based Method for Efficient and Highly Specific Screening of Genome Edited Cells. *PLoS ONE*, 11(4), e0153901–17. doi:10.1371/journal.pone.0153901

Curriculum Vitae

Scott D. Findlay

Education

Western University: London, Ontario, Canada

PhD candidate (2016), September 2010 – December 2016

Department of Anatomy and Cell Biology and the Collaborative Graduate Program in Developmental Biology, the Schulich School of Medicine and Dentistry

The University of Western Ontario: London, Ontario, Canada

Bachelor of Medical Sciences (BMSc)

Honors Specialization in the Biochemistry of Infection and Immunity

Scholar's Electives Program

Graduated with distinction, June 2010

Scholarships held

Ontario Graduate Scholarship (OGS). $15,000 for 2014-2015.

Natural Sciences and Engineering Research Council of Canada (NSERC) Postgraduate Scholarship D—3 years (PGS D3). $63,000 over 36 months. 2011-2014.

Natural Sciences and Engineering Research Council of Canada (NSERC) Alexander Graham Bell Canada Graduate Scholarship (CGSM). $17,500 over 12 months. 2010-2011.

Scholarships Awarded and Declined

Ontario Graduate Scholarship (OGS), with distinction. $15,000 for 2013-2014.

Ontario Graduate Scholarship (OGS). $15,000 for 2011-2012.

Ontario Graduate Scholarship (OGS). $15,000 for 2010-2011.

Awards and Honours

"Best posters" winner at Cancer Research Institute of Northern Alberta (CRINA) Research Day. Edmonton, Alberta, November 2015.

"Best Overall Oral Presentation" winner at RiboWest 2015. Edmonton, Alberta. June 2015.

Oral presentation award. University of Alberta Signal Transduction Research Group Annual Retreat, Edmonton, Alberta. May 2015.

Poster award. CIHR CaRTT Oncology Research and Education Day, London, Ontario. June 2013.

Selected to attend the Canadian Student Health Research Forum as a top 10% PhD candidate in the Schulich School of Medicine and Dentistry, June 2013, Winnipeg, Manitoba.

Poster award: Honorable mention (3rd place). Canadian Student Health Research Forum. Winnipeg, Manitoba. June, 2013.

Accepted to the Collaborative Graduate Program in Developmental Biology, September 2010.

University of Western Ontario Developmental Biology Student Award Winner. $1,250. August 2010.

Western Gold Medal - Honors Specialization in Biochemistry of Infection & Immunity (Awarded to the graduating student with the highest average in the honors specialization), June 2010.

Dean's Honor List, 2009-2010.

Dean's Honor List, 2008-2009.

UWO In-Course Scholarship. $700, 2008.

Dean's Honor List, 2007-2008.

Dean's Honor List, 2006-2007.

Western Scholarship of Excellence, 2006.

Academic Employment History

Research Assistant

Dr. Paul Thagard, University of Waterloo: Waterloo, Ontario

May 2010- August 2010

May 2009- August 2009

May 2008- August 2008

Publications:

Findlay S.D.*, Vincent K.M.*, Berman, J.R., & Postovit L.M. (2016). A Digital PCR-Based Method for Efficient and Highly Specific Screening of Genome Edited Cells. *PLoS ONE* 11(4): e0153901. doi: 10.1371/journal.pone.0153901.

Findlay, S.D. & Postovit. L.M. (2016). Common Genetic Variation in Chromosome 10 q22.1 Shows a Strong Sex Bias in Human Embryonic Stem Cell Lines and Directly Controls the Novel Alternative Splicing of Human NODAL which is Associated with XIST Expression in Female Cell Lines. *Stem Cells,* 34(3):791–796.

Vincent, K.M, **Findlay S.D.**, & Postovit L.M. (2015). Assessing breast cancer cell lines as tumour models by comparison of mRNA expression profiles. *Breast Cancer Research,* 17:114.

Findlay, S.D. & Thagard, P. (2014). Emotional change in international negotiation: Analyzing the Camp David accords using cognitive-affective maps. *Group Decision and Negotiation,* 23: 1281-1300.

Quail, D.F., Zhang G., **Findlay S.D.**, Hess D.A., & Postovit L.M. (2014). Nodal promotes invasive phenotypes via a non-canonical Mitogen Activated Protein Kinase-dependent pathway. *Oncogene.* 33(4):461-73.

Findlay, S.D. & Thagard, P. (2012). How parts make up wholes. *Frontiers in Physiology*, 3: 455.

Quail, D.F., Zhang G., Walsh L.A.,Siegers G.M., Dieters-Castator D.Z., **Findlay, S.D.**, Broughton H., Putman D.M., Hess D.A. & Postovit L.M. (2012). Embryonic morphogen Nodal promotes breast cancer growth and progression. *PLoS One*, 7(11): e48237.

Findlay, S.D.*, Jewer M*. & Postovit, L.M. (2012). Post-transcriptional regulation in cancer progression: Microenvironmental control of alternative splicing and translation. *Journal of Cell Communication and Signaling,* 6(4):233-48.

Quail, D.F., Walsh L.A., Zhang G., **Findlay S.D.**, Moreno J., Fung L., Ablack A., Lewis J.D., Done S.J., Hess D.A., & Postovit L.M. (2012). Embryonic Protein Nodal Promotes Breast Cancer Vascularization. *Cancer Research*, 72:3851-3863.

Thagard, P. & **Findlay, S.** (2011). Conceptual Change in Medicine: Explaining Mental Illness. In W. J. Gonzalez (ed), Conceptual Revolutions: From Cognitive Science to Medicine, Netbiblo, A Coruña.

Thagard, P. & **Findlay, S. D.** (2011). Changing minds about climate change: Belief revision, coherence, and emotion. In E. J. Olsson & S. Enqvist (Eds.), Belief revision meets philosophy of science (pp. 329-345). Berlin: Springer.

Thagard, P. & **Findlay, S.** (2010). Getting to Darwin: Obstacles to Accepting Evolution by Natural Selection. *Science & Education*, *19*: 625-636.

*Authors contributed equally to the work.

External Presentations:

Selected abstracts for oral presentation:

"Alternative splicing of a novel human NODAL transcript"

Scott Findlay and Lynne-Marie Postovit

RiboWest Meeting, Edmonton, Alberta, June 2015.

"Alternative splicing of a novel human NODAL transcript"

Scott Findlay and Lynne-Marie Postovit

The Canadian Cancer Research Conference, Toronto, Ontario, November 2013.

"Alternative splicing of a novel NODAL transcript in human pluripotent stem cells"

Scott Findlay and Lynne-Marie Postovit

Till & McCulloch Meetings, Banff, Alberta, October 2013.