Lineage specification of pluripotent populations in murine development

by

Brian DeVeale

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy

Graduate Department of Molecular Genetics in the University of Toronto

© Copyright by Brian DeVeale (2013)
Abstract

“The scientist, by the very nature of his commitment, creates more and more questions, never fewer. Indeed the measure of our intellectual maturity, one philosopher suggests, is our capacity to feel less and less satisfied with our answers to better problems.”

~G.W. Allport, Becoming, 1955

It will be interesting to look back at this thesis in a few decades and reflect on how the questions and interpretation of data in the field of developmental biology have changed. Indeed, a biologist currently in their twilight years might reflect on their youth, before the discovery of hereditary material, and compare that bookend with the range of genome sequences and related knowledge currently available. How long will it take before this thesis reads like a debate about whether the male or female contributed the ‘homunculus,’ a miniature preformed human to the embryo that grows into an adult?

In this thesis I asked three related questions: whether the role of Oct4 during embryogenesis provides insight into its contribution to pluripotency; how surfaceome changes contribute to functional maturation of neural stem cells and to what extent the murine genome is imprinted. Our data indicate that Oct4 is required for posterior expansion. We propose that the function of the protein is conserved, but that its expression has been coopted to yield different cell types based on its combination with different factors. We show that fundamental aspects of cell biology are altered during the maturation from pluripotent populations to neural stem cells, and identify mediators of proliferation, survival and adhesion that distinguish neural stem cell regulation from their precursors. Finally, we validated discovery of a dozen novel imprinted transcripts using a genomic approach. These discoveries will contribute to a holistic view of the
causes and consequences of imprinting, but do not support a paradigm shift in the scale and consequences of imprinting.
Acknowledgements

Graduate school has been a unique educational experience for me. The cliché of learning more from my failures than successes certainly held true, and I would extend that to say that I learned larger lessons from larger failures. I would say that I am a better scientist than when I started, in large part thanks to my supervisor Derek van der Kooy. I thank him for that, as well as his mentorship over the duration of a long thesis. Everyone and their cousin preach the merit of critical thinking, and I would say that I have developed considerably along this spectrum because of his approach to instilling it. This lesson stands out most, but I would also thank him for giving me the freedom to dig myself out of holes, his scientific insight, his enthusiasm, and emphasis on effective communication. Finally, his humour has made the tough times tolerable: “geneticists from Texas can’t be taken seriously,” “I’m surprised they aren’t ashamed to put their names on that,” and the two-handed Nixon gesture whenever he thought he won a bet.

I would like to thank everyone else who was involved in the scientific portion of my graduate studies. To my committee members, notably Freda Miller who lasted the entire marathon, as well as Jane Aubin, Peter Zandstra and Brian Ciruna, for their supportive approach, insightful comments, difficult questions and the length of rope they gave me to work with. Thanks to major collaborators in Damaris Bausch-Fluck and Bernd Wollscheid on the surfaceome project, as well as Lee Lim and Tomas Babak on the imprinting work – I learned a lot from each and had a lot of fun doing so. I also received tremendous technical support and camaraderie from Brenda Coles, Susan Runciman, AJ Wang, Rick Stevens, Phil Garrett-Engele and Marina Gertsenstein over the years. I would probably have a lot less hair and data to show for my time in the lab if it was not for their help.

I would like to thank other investigators for sharing resources. To Andras Nagy for Z/EG mice, R1 ES cells and RFP ES cells; to Alexey Tomilin and Hans Scholer for providing the
Oct4\textsuperscript{fl} mice; to Larysa Pevny for providing the Sox2\textsuperscript{fl} mice; to Achim Gossler for the Bry-Cre mice; to Shinichi Nishikawa for the Sox1-Cre mice; to Deborah Chapman for the Dll1-Cre mice; to Hitoshi Niwa for ZHBTc4 and ZHBTc4 ES lines; to Robin Lovell-Badge for his Sox1 antibody; Sean Wu for Nkx2.5 ES cells and to Tania Petruzzello and Phil Marsden for castaneous mice on short notice.

I would also like to thank my peers who have been great teachers and friends over the years. For the first part of my degree, Phil Karpowicz, Drew Heinmiller, Radha Chaddah, Simon Smukler, Hector Vargaz-Perez, Laura Clarke, Anthony Lin, Taryn Grieder and Christina Thiele were a great team to hang out with. More recently, Kento Onishi, Emmanuel Nazareth, Charles Yoon, Elizabeth Csaszar, Shreya Shukla, Mukul Tewary, Weijia Wang, Lily Allen, Curtis Woodford, Hannah Song, Nims, Petra and Brian Ballios. Also, a special thanks to Jessica Sturgess for a lot of great memories.

Finally, I owe an enormous debt to the most important people in my life whom I’m lucky to have. To all my other friends who want nothing to do with this thesis and have always been a great social outlet. To my family who has been unwavering in their support and fiercely loyal in all situations. To my sister and lifelong best-friend, to my mother for her unconditional love and to my father for being a great mentor. And to my spicy bean – Jessica Millar. I love you deeply, and thank you for opening my eyes, pushing my limits, and being the perfect complement. I hope we share a lifetime of happiness.
Table of Contents

List of Tables and Figures pg viii

List of Abbreviations pg xi

Chapter 1: Introduction pg 1
The origins of multicellular life pg 1
Molecular origins of multicellular life pg 3
Molecular origins of complexity in multicellular species pg 4
Multicellular development programs pg 6
Molecular basis of basal body plan establishment: axis formation pg 13
Establishment of a consistent body plan: lineage specification pg 14
Establishment of a consistent body plan: spatial coordination of lineages pg 17
Evolution of differing body plans pg 19
Setting the balance between mosaic and regulative modes of development pg 20
Pluripotency in mice pg 23
Tradeoffs in selective pressures defining lineages and body plans pg 27
The influence of sexual reproduction on body plans pg 27
The influence of intrinsic cellular constraints on body plans pg 29
Evolution of the Nervous System pg 31
Contribution of neuroepithelial organization to cortical evolution pg 33
Division mode in the SVZ and its relation to cortical expansion pg 35
Adult neurogenesis pg 35
Evolution of the germ line from somatic SC populations pg 37
Imprinting pg 40

Chapter 2: Oct4 facilitates the expansion of murine posterior progenitors after E7.0.
Abstract pg 46
Introduction pg 47
Materials and Methods pg 51
Results pg 54
Discussion pg 79

Chapter 3: Surfaceome profiling reveals novel regulators of neural stem cell function.
Abstract pg 86
Introduction pg 86
Materials and Methods pg 88
Results pg 94
Discussion pg 111

Chapter 4: Global survey of genomic imprinting by transcriptome sequencing.
Abstract pg 116
Introduction
Materials and Methods
Results
Discussion

Chapter 5: General Discussion
Oct4 facilitates the expansion of murine posterior progenitors after E7.0.
The molecular mechanism of pluripotency
Alternatives to a levels based model
Resolution of the alternative pluripotency models
Understanding Oct4’s mechanistic contribution to pluripotency
Cellular potency in the context of development
Evolutionary differences that might be expected between mice and humans

Surfaceome profiling reveals novel regulators of neural stem cell function.
Relevance of the proteins implicated in NSC function
Utility of this analysis
Evolutionary differences that might be expected between mice and humans

Global survey of genomic imprinting by transcriptome sequencing.
Improved prediction accuracy will make RNA-seq a sensitive imprint
discovery tool
Towards the underlying cause of imprinting: spatial, temporal and
phylogenetic studies
Towards the underlying cause of imprinting: revisiting function
Evolutionary differences that might be expected between mice and humans

References
List of Tables and Figures

Chapter 1: Introduction
Figure 1. The evolution of multicellular organisms from unicellular Choanoflagellates. pg 9

Chapter 2: Oct4 facilitates the expansion of murine posterior progenitors after E7.0.
Figure 1. Oct4 localization from E6.5-9.25 of murine development. pg 48
Figure 2. Kinetics of Oct4 depletion following conditional removal pg 55
Figure 3. The Oct4<sup>COND MUT</sup> phenotype. pg 56
Figure 4. Characterization of the Oct4<sup>COND MUT</sup> phenotype. pg 58
Figure 5. Characterization of the Oct4<sup>COND MUT</sup> phenotype continued pg 60
Figure 6. Timing of Oct4 depletion is relevant to penetrance of the Oct4<sup>COND MUT</sup> phenotype, but diversion of cells between germ layers is not. pg 64
Figure 7. Oct4 is required in the embryo after E7.0, but not in extraembryonic tissue pg 66
Figure 8. Depletion of core pluripotency network factor Sox2 does not phenocopy Oct4 depletion. pg 69
Figure 9. c-Myc, Smad1 and Oct4 targets are up-regulated while targets co-occupied by Oct4 and Sox2 are down-regulated coincident with Oct4 depletion. pg 70
Figure 10. Differential gene expression coincident with loss of Oct4. pg 74
Figure 11. Decreased proliferation contributes to Oct4<sup>COND MUT</sup> phenotype pg 76
Figure 12. Cdx2, p-Smad1 and Chordin domains are intact in Oct4<sup>COND MUT</sup>, as is early lineage specification of mesoderm and neuroepithelium. pg 78

Chapter 3: Surfaceome profiling reveals novel regulators of neural stem cell function.
Figure 1. Overview of CSC-based protein identifications. pg 97
Table 1. Cell surface proteins showing the greatest fold enrichment in pNSCs relative to ES cells pg 99
Table 2. Cell surface proteins showing the greatest fold enrichment in dNSCs relative to ES cells pg 99
Table 3. Cell surface proteins showing the greatest fold enrichment in dNSCs relative to pNSCs

Figure 2. NSCs lose the ability to colonize the ICM as they mature through the lineage.

Figure 3. Lineage distribution bias of pNSC in chimeras.

Figure 4. pNSC distribution is not the result of fusion.

Figure 5. Upregulation of N-cad prevents dNSCs from aggregating with morula

Figure 6. The ES-derived neural lineage predicts molecules that are functionally relevant for NSCs derived from adult mice.

Figure 7. Independent confirmation that inhibition of the molecules of interest is responsible for the effects on NSC function.

Figure 8. Schematic summarizing the neural maturation model

Chapter 4: Global survey of genomic imprinting by transcriptome sequencing.

Table 1. Primer pairs used in cDNA-RFLP and RT-PCR Sanger confirmations.

Table 2. Primers used to validate or refute putative imprinted loci by pyro-sequencing.

Figure 1. Expression bias measured by pyro-sequencing is reproducible in biological replicates.

Figure 2. Calibrating SNP discovery.

Figure 3. Measuring allelic expression by detecting and counting SNPs.

Figure 4. Genome-wide identification of imprinted regions.

Figure 5. Allele-biased expression of known transcripts.

Table 3. Summary of validated novel features.

Figure 6. Imprinting beyond documented transcriptional boundaries.

Figure 7. Independent replication with E17.5 brains.
Figure 8. False discoveries explain the majority of novel imprinted genes. pg 142

Figure 9. False-discoveries explain the majority of sex-specific imprinted genes. pg 144

Figure 10. The majority of sex-specific imprints reported are due to technical and biological variance. pg 145

Figure 11. Validating features predictive of genomic imprinting. pg 147

Figure 12. Genomic clustering of candidate imprinted transcripts. pg 148

Figure 13. Pyro-sequencing traces confirmed 6 of the 11 predicted novel parent-of-origin effects in E17.5 brains. pg 150

Figure 14. Concordance in regional imprinting for 53 novel imprinted genes pg 156
List of Abbreviations

Antp  Antennapedia
A/P  anterior/posterior
Ascl1  Achaete-scute complex homolog 1
ASE  allele-specific expression
Aspm  Abnormal spindle-like microcephaly-associated protein
ATI  after tamoxifen induction
B  C57BL6J
Bax  Bcl-associated X protein
β-catenin  β-cadherin associated protein
Bcl2  B cell leukemia/lymphoma 2
BDNF  Brain-derived Neurotrophic Factor
BMP  bone morphogenetic protein
Bry  Brachyury
C  CAST/EiJ
Cdkn1c  Cyclin-dependent kinase inhibitor 1c
Cdx2  Caudal type homeobox 2
CSC  Cell Surface Capture
DE  distal enhancer
DLHPs  dorsolateral hinge points
Dll1  Delta-like 1
DMRs  differentially methylated regions
dNSCs  definitive NSC
dpc  days post coitum
E-cad  E-cadherin
Efnb3  Ephrin B3
EGF  Epidermal Growth Factor
Eomes  Eomesodermin
EphA4  Ephrin A4
ES  embryonic stem
FCS  fetal calf serum
FDR  false-discovery rate
FGF  Fibroblast growth factor
Gata  Gata-binding
GFAP  glial fibrillary acidic protein
HhIP  hedgehog-interacting protein
i-cells  interstitial cells
ICM  inner cell mass
ID  inhibitors of differentiation
Igf2  Insulin growth factor 2
Insc  Inscuteable
IS  imprinting score
Klf  Kruppel-like factor
LC MS/MS  liquid chromatography coupled to tandem MS
LIF  Leukemia inhibitory factor
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHP</td>
<td>median hinge point</td>
</tr>
<tr>
<td>MS</td>
<td>mass-spectrometry</td>
</tr>
<tr>
<td>NaIO4</td>
<td>sodium meta-periodate</td>
</tr>
<tr>
<td>Nanog</td>
<td>Nanog homeobox</td>
</tr>
<tr>
<td>Nb</td>
<td>neuroblasts</td>
</tr>
<tr>
<td>N-Cad</td>
<td>Neural-cadherin</td>
</tr>
<tr>
<td>ncRNAs</td>
<td>non-coding RNAs</td>
</tr>
<tr>
<td>NE</td>
<td>neuroepithelium</td>
</tr>
<tr>
<td>NGS</td>
<td>normal goat serum</td>
</tr>
<tr>
<td>NSCs</td>
<td>neural stem cells</td>
</tr>
<tr>
<td>NTC</td>
<td>neural tube closure</td>
</tr>
<tr>
<td>NXS/T</td>
<td>asparagine, any amino acid, serine or threonine</td>
</tr>
<tr>
<td>PCP</td>
<td>planar cell polarity</td>
</tr>
<tr>
<td>PE</td>
<td>proximal enhancer</td>
</tr>
<tr>
<td>PFC</td>
<td>medial prefrontal cortex</td>
</tr>
<tr>
<td>PGCs</td>
<td>primordial germ cells</td>
</tr>
<tr>
<td>PH3</td>
<td>Phospho-histone H3</td>
</tr>
<tr>
<td>piRNAs</td>
<td>Piwi-interacting RNAs</td>
</tr>
<tr>
<td>pNSCs</td>
<td>primitive NSCs</td>
</tr>
<tr>
<td>POA</td>
<td>preoptic area</td>
</tr>
<tr>
<td>PRC2</td>
<td>polycomb repressive complex 2</td>
</tr>
<tr>
<td>p-Smad1</td>
<td>phosphorylated-Smad1</td>
</tr>
<tr>
<td>PVZ</td>
<td>periventricular zone</td>
</tr>
<tr>
<td>RFP</td>
<td>red fluorescent protein</td>
</tr>
<tr>
<td>RTK</td>
<td>receptor tyrosine kinases</td>
</tr>
<tr>
<td>SC</td>
<td>stem cell</td>
</tr>
<tr>
<td>SE</td>
<td>subependyma</td>
</tr>
<tr>
<td>SGZ</td>
<td>supragranular-layer zone</td>
</tr>
<tr>
<td>Shh</td>
<td>sonic hedgehog</td>
</tr>
<tr>
<td>Smo</td>
<td>Smoothened</td>
</tr>
<tr>
<td>SNPs</td>
<td>single nucleotide polymorphisms</td>
</tr>
<tr>
<td>Sox2</td>
<td>Sry-box containing gene 2</td>
</tr>
<tr>
<td>Stat3</td>
<td>Signal Transducer and Activator of Transcription 3</td>
</tr>
<tr>
<td>SVZ</td>
<td>subventricular zone</td>
</tr>
<tr>
<td>Tbx6</td>
<td>T-box 6</td>
</tr>
<tr>
<td>TE</td>
<td>trophectoderm</td>
</tr>
<tr>
<td>TFs</td>
<td>transcription factors</td>
</tr>
<tr>
<td>TGF</td>
<td>Transforming growth factor</td>
</tr>
<tr>
<td>TPP</td>
<td>Trans-Proteomic Pipeline v4.3</td>
</tr>
<tr>
<td>UTRs</td>
<td>untranslated regions</td>
</tr>
<tr>
<td>Wnt</td>
<td>Wingless-related</td>
</tr>
<tr>
<td>WT</td>
<td>wild-type</td>
</tr>
</tbody>
</table>
Introduction

Ongoing adaptation enabling expansion offers a rare biological law: “...each slight variation, if useful, is preserved.” This principle of Natural Selection manifests as ongoing specialization – that processes relevant to survival and procreation become increasingly efficient. Specialized cellular lineages offer a striking example of the capacity selective pressure has to mould efficient systems. These lineages provide multicellular organisms with an array of useful features: protection from desiccation, enhanced nutrient extraction and storage, mobility, learning and memory, communication, sensation, thermoregulation, detoxification and immunity from pathogens. Indeed, the utility of specialized lineages is evident from the distinct occasions that multicellular modes of life evolved [1, 2]. Myxobacteria is a prokaryotic example, while plants, animals, fungi, green algae, brown algae and slime moulds are separate eukaryotic instances [3].

However the principle of Natural Selection does not provide a sequential account of the selective pressures shaping each species or the mechanisms by which useful variations are preserved. Delineating evolutionary histories enables layering of detail from progressively more derived organisms to dissect intricate aspects of biology. In this thesis, I asked specific questions concerning the role of pluripotency factors during development, the role of cell surface molecules during lineage specification and unconventional modes of inheritance, so I will start by providing evolutionary history as context for these projects.

The origins of multicellular life

Species diversity ultimately reflects a distinct series of selective pressures from a common ancestor, referred to as the urmetazoan [3]. As such, the origins of multicellular life provide a foundation for understanding highly derived aspects of complex multicellular organisms. While the majority of the history of life on earth --the first 3.5 million years--
involved only unicellular organisms, multicellularity evolved ~600 million years ago, just prior to the Cambrian explosion [4]. Several independent parallel evolutions of multicellular life suggest considerable advantage over unicellular approaches, including size to escape predation [5] as well as specialized tissues to optimize and spatiotemporally separate processes [6].

There are two modes by which multicellular life arose: cell-aggregation and non-division. Cell-aggregation occurred terrestrially, applies to organisms such as slime moulds, and as the name suggests indicates that multicellular life arose by aggregation of unicellular organisms [3]. Conversely, non-division occurred aquatically, applies to organisms such as plants, animals and fungi, and indicates that multicellular life arose when two daughters of a cell division were kept in close proximity [3]. Plants and fungi can be further subdivided from animals on the basis that cell mobility within them is restricted by a cell wall or other matrix: daughter cells remain in the same position relative to each other. Conversely the relative position of daughter cells is plastic in animals.

The genetic program for multicellular life evolved through gradual cooption of content from unicellular genomes. Independent evolutions of unicellular organisms to multicellular modes of life are supported by efforts combining comparative genomics with phylogeny (phylogenetics). This approach revealed that all of the tools required for multicellular life are present in unicellular organisms. Receptor tyrosine kinases (RTK), immunoglobulins and Sox-family transcription factors (TFs) that once were considered animal specific are in fact present in the unicellular choanoflagellate M. brevicollis [1, 7-9]. Superficially, it is curious that proteins which contribute to processes required by multicellular organisms are found in unicellular organisms where these functions are irrelevant, however these observations are reconciled by the model that these genes are coopted to serve different functions during evolution. To illustrate,
homeodomain TFs that pattern multicellular body plans regulate the haploid-diploid transition in the unicellular alga *Chlamydomonas* [10], and immunoglobulin domains likely contributed to adhesion as they were found within cadherins proteins [11].

**Molecular origins of multicellular life**

Multicellular life was established by cooption of genes that provided immediate utility, beginning with cell-cell adhesion. The core component of animal adhesion junctions are cadherins. Although these were once thought to be animal specific, a considerable repertoire is present in unicellular organisms. By way of example, there are 23 in unicellular *M. brevicollis* [11]. This is a larger set than is present in the relatively complex invertebrate *D. melanogaster* [11]. ‘Complexity,’ as it will throughout, refers to the number of functionally distinct cell types per organism. Cadherins are speculated to have had a role in sensing and responding to external nutrient cues, or capture of bacterial prey in the feeding structure of *M. brevicollis* before cooption to a role in cell-cell adhesion [11].

Integrins and septate junctions may have been present with adhesion junctions in ancestral multicellular species. Like cadherins, integrins are found in unicellular relatives of animals such as *C. owczarzaki* [12]. Integrins contribute to hemidesmosomes and focal adhesions that adhere to extracellular matrices, enabling cell polarity and barrier formation as epithelium. Septate junctions also contribute to epithelial solute barriers that arose in ancestral sponges [13-16], supporting a role early in the evolution of multicellular life.

The correlation of gap junctions with cell-cell communication and protocadherins with intricate organization of more complex organisms suggests enabling roles in evolution of these respective traits. The dynamic nature of relevant functions is highlighted by this evolutionary sequence: the utility of cell-cell communication and intricate organizational patterns would only
have arisen after cell-cell adhesion enabled the transition to multicellular life. Consistent with this, the most ancestral gap junctions are found in cnidarians such as sea anemones and hydra [17-19]; these also have adherens and septate junctions and are derived relative to basal sponges.

**Molecular origins of complexity in multicellular species**

Cooption of genes in unicellular organisms to new roles in multicellular organisms accounts for only a fraction of organism diversity. Phylogenetics suggests that *de novo* functions have been created through alteration in regulatory elements [20, 21], cooption of duplicated genes [22], and shuffling of genetic modules [9].

Changes in the developmental timing, distribution or abundance of a gene can have functional consequences [20, 21]. Increases in species complexity correlate better with increases in intergenic DNA than increases in gene number [22]. Placed in perspective, 30 genes occupy the same length of DNA in the human genome as an average prokaryotic genome which contains thousands of genes [22]. The genomes of *S. cerevisiae, C. elegans, D. melanogaster, M. musculus* and *H. sapien* contain ~6000, ~20 000, ~15 000, ~24 000 and ~24 000 genes respectively, and are ~12 Mb, ~100 Mb, ~140 Mb, ~3000 Mb and ~3000 Mb, respectively. This increase in intergenic DNA is likely not entirely functional (nearly 50% of the human genome is derived from transposable elements [23]), but the increase in transcriptional regulation contributed by this intergenic DNA does correlate with species complexity. Increases in the prevalence of non-coding RNA (ncRNA), alternative splicing [24, 25], untranslated regions (UTRs) as well as cis-acting elements all contribute. ncRNAs that mediate post-transcriptional regulation of coding transcripts became fixed at the origin of vertebrate evolution, being absent in non-chordates and present in lampreys [26]. Alternative splicing contributes to tissue-specific gene function [27]. UTRs enhance precision in the abundance, localization and kinetics of
translation [28]. While cis-acting elements of related genes can be edited to alter throughput of a multi-component process: increases or decreases that maintain stoichiometry of pathway components support coordinate positive selection of adaptive regulatory variants [29]. Cis-acting variations in single genes might alter the timing or distribution of its expression leading to alterations in networks by producing distinct gene combinations and pleiotropic effects [20, 21, 30]. While systematic genomic analysis is required to settle the debate, regulatory variants –as opposed to coding changes-- are proposed to account for the majority of adaptive changes [31-33].

Gene duplication is relatively common, and makes a major contribution to organism evolution at the molecular level; estimates vary according to methodology, but gene duplication events are believed to be similar or higher in frequency (1000x at the high end of the estimate) to the rate at which base substitution occurs per nucleotide [34]. This translates to estimates ranging from $1 \times 10^{-3}$-$1 \times 10^{-9}$/gene/generation depending on the methodology and organism used for estimation [34]. To illustrate the consequences numerically, the mammalian genome has ~1500 TFs compared to ~600 putative TFs in C. intestinalis [35]. The identity of the TF families that expanded varies by kingdom: homeodomain proteins that have a major role in patterning animals [36], are also found in fungi and plants but serve a relatively minor role in their patterning. Conversely, the MADS-box family is expanded and have major roles in patterning plants are also found in animals, fungi and other eukaryotes, but have narrow roles in their development [36, 37]. Gene duplications are presumed to be generated primarily by non-allelic homologous recombination between transposable elements at different loci [38]. Following duplication, one copy of a gene remains under tight selective pressure, while functional constraint on the other
copy is relieved: it may be coopted or more commonly decay. Adaptive selection following loss of constraint is a recurring principal driving evolution at various scales.

Alternative combinations of existing genetic modules can also alter gene-function. For example, domain composition and organization can effect interactions with substrates, ligands or co-factors. Hedgehog is one example proposed to have evolved from ‘domain-shuffling’ early in bilaterian animal evolution. Choanoflagellates have the ‘hog’ domain [17-19], while the ‘hedge’ domain is first observed in early-branching animals such as sponges and cnidarians [17-19]. More TFs acquired domains through shuffling during animal than plant evolution, such that there are a greater number of animal-specific TFs, and metazoans have the greatest diversity of protein domains [3]. Conversely, expansion by each protein domain that did so during plant evolution is more extensive.

Multicellular development programs

Positive selection manifests at different scales. Single genomes were derived to yield the programs for multiple specialized cell types and complex body plans. At the base of the animal phylogenetic tree, sponges provide an entry point to the origins of cellular lineage and body plan diversity.

Compared to the stereotyped body plans of familiar animals, sponge development and adult body plans are flexible and responsive to external signals. The sponge life-cycle is generally described as development from a zygote to an adult before sexual reproduction initiates a new cycle, but sponges also bud off asexually as ‘gemmules’ [39, 40]. Gemmule formation is not part of an intrinsic cycle but induced by environmental factors such as temperature or drought [39]. Lineage specific gene-expression programs are similarly responsive to environmental cues: upregulation of the epithelial program occurs in response to water currents.
and of the skeletal program in response to silicon and ferric iron [39, 41]. Sponges also are plastic at higher levels of organization; they routinely remodel their canal systems, reconstruct tissues after sex, and connect new tissue to the existing canal networks [42].

Sponge plasticity is dependent on two stem cell (SC) populations. Sponges dissociated into single cells will aggregate to form ‘primmorphs,’ a contrived developmental stage that produces a mature sponge. However if these aggregates are depleted of a SC population referred to as archeocytes, they fail to mature [39]. The inability of the remaining somatic lineages to pattern a mature body plan, produce gametes or produce the somatic lineages required after aggregation supports at least two tiers of cellular potency in sponges. The second SC population are called choanocytes, which self-renew and produce sperm. Archeocytes produce choanocytes, somatic lineages, oocytes as well as thesocytes that comprise the germinal portion of gemmules. Thus these two SC populations together are required and sufficient for sexual reproduction, while archeocytes are required and sufficient for asexual reproduction. So how did the transition from unicellular choanoflagellates to these multicellular sponges occur?

The similarity of choanocytes to unicellular choanoflagellates forms the basis of one model (derived from [43], [1] and [42]) (Fig 1). The model posits that the first multicellular organism was a spherical conglomerate of adherent choanoflagellates with external flagellum for feeding and motility, similar to the larval stage of sponge development (Fig 1B). Choanocytes and choanoflagellates are functionally similar in that both self-renew and absorb nutrients: choanocytes line sponge chambers, absorb nutrients and pass these to other cells via vesicles. Morphologically the two are similar in that they both possess a single flagellum surrounded by microvilli. Choanocytes create currents within sponge chambers using flagella --that propel choanoflagellates-- and sift through the content of this flowing water. The crux of this model
rests on whether choanocytes interconvert with archeocytes, and this point is debated [42]. Choanocytes typically are described as differentiated cells, however it also is reported that they are a stock for archeocytes, generating them prior to gemmule formation or during remodelling [42, 44, 45].

The model proposes that choanocyte internalization from the spherical conglomerate relieved their functional constraint of nutrient intake and the associated morphological constraint of polarity (Fig 1C). Without the benefit of each intermediate the precise sequence by which cellular lineages evolved in sponges cannot be determined,
Figure 1. The evolution of multicellular organisms from unicellular Choanoflagellates. A Unicellular Choanoflagellates B Conglomerate of Choanoflagellates C Choanocytes evolve from Choanoflagellates in the first stage of sponge evolution. Loss of polarized morphology and internalization of a Choanocyte yields the second stem cell population -- Archeocytes. D Evolution of Thesocytes from Archeocytes may have occurred prior to evolution of somatic lineages. Thesocytes enable asexual reproduction during times of environmental stress. Archeocytes produces clusters of Thesocytes called gemmules that are released in response to environmental stress. E Axis formation in gemmules is initiated by external cues. The first stage of body plan establishment, oscule formation, is initiated by contact with a solid surface. F Sexual reproduction in sponges involves Choanocytes generated by morphologically similar Choanocytes, whilst Oocytes evolved from Archeocytes. Sponge development following sexual reproduction precedes through the conventional morula, blastula, and gastrula stages before a mature sponge is formed. Sponges formed via asexual and sexual reproduction are equivalent. G Proportions and activity of somatic lineages are environmentally influenced. Silicon and iron induce formation of Sclerocytes their production of skeletons.

however lineage relationships offer insight. Epithelialisation to enhance nutrient intake would seem favourable in choanocyte/choanoflagellate conglomerates, and suggest the intuitively attractive possibility that epithelium arose from choanocytes, however the production of sponge epithelium from archeocytes supports its evolution from archeocytes. Similarly, lineage relationships suggest that thesocytes, oocytes and somatic lineages arose from archeocytes in spite of the benefits that likely would have been provided by structural rigidity and stress resistant buds had they arisen directly from epithelium (Fig D-G). These lineage relationships support a model where loss of constraint accompanied internalization of choanocytes, which in turn facilitated cooption of this population for a variety of uses [41].

Since natural selection is predicated on immediate utility, archeocytes must have provided choanocyte conglomerates a reproductive or survival advantage that exceeded the benefit of nutrient absorption. One possibility is that their utility was comparable to stress-resistant spores found in unicellular bacteria and yeast. In this scenario archeocytes would be the first stage in the evolution of germinal thesocytes that are induced by environmental stresses and
optimized to endure them (Fig 1C,D). Cells that are not required during favourable growth conditions would be liberated to evolve optimal stress resistance. One key difference from unicellular organisms would be that archeocytes maintained contact with choanocyte conglomerates. An alternative is that archeocytes were not bystanders during periods of environmental richness, but behaved as parasites, perhaps optimized to colonize the nutrient rich interior of neighbouring choanocyte conglomerates. There is precedent for this: the vasculature of colonial ascidians *B. schlosseri* fuses with neighbours providing the opportunity for germline parasitism by genetically distinct SC neighbours [46]. Fusion only occurs between individuals that share a histocompatibility allele, suggesting that this allele arose to protect the *B. schlosseri* germline from parasitism by neighbours [46]. Population expansion is facilitated in both scenarios, during stressful and favourable conditions respectively.

Establishment of a second cell type was the first step towards evolution of the modern sponge body plan that develops by one of two routes in nature. Following fertilization, sponge development begins as it does in other metazoans with cleavage divisions to a morula, blastula and possibly gastrula before embryos develop into mature larvae, become sessile and establish the adult body plan [39, 47] (Fig 1F). Alternatively sponges can develop from gemmules that do not progress through these stereotyped cleavage divisions, beginning instead as bundles of thousands of thesocytes [42], but ultimately reaching the same endpoint (Fig 1D,E). Meiotic genes in the Choanoflagellate genome led to speculation that there may be a sexual phase to their life-cycle, however it does not appear constitutive: Choanoflagellates primarily reproduce asexually [48], suggesting that sponge reproduction was initially asexual, and that this mode of reproduction might be a more accurate reflection of how their patterning and body plans arose.
Development is typically described as a stereotyped series of events where cell division, fate and position are coordinated, however asexual sponge reproduction would obviate the need for sequential evolution, as the sequential series of events required following sexual reproduction are not necessarily required for asexual development. The plasticity required to develop a mature body plan from a mass of cells suggests that mechanisms are present to sense and coordinate cellular needs and their positional requirements. Body plans may have evolved through regulatory responses to the environment and paracrine signals from other members of a conglomerate. The sponge organizer and primary body axis are established perpendicular to external surfaces contacted by sessile larvae (Fig 1E). This contact breaks symmetry and initiates establishment of the body plan. Attachment to a substratum triggers morphogenesis, indicating incorporation of environmental signals into layout of the body plan [39, 42]. Similarly, sponge morphogens, and ultimately progeny of the fate they promote are induced by environmental stimuli: water currents lead to induction of Iroquois associated with the pinacocyte or epithelial layer, while orthologues of Mesenchymal Stem Cell-like Protein and Noggin are induced by the inorganic elements silicon and ferric iron and associated with sclerocyte or skeleton producing cells [39] (Fig 1G). Considerable environmental feedback is incorporated into sponge maturation, with lineage proportions being environmentally regulated and adults varying in size according to age and environmental richness [49]. Sponges are not an exception in having a flexible body pattern and development program that is tailored by environmental cues. Using a dramatic illustration, sea urchins and tunicates have extremely flexible developmental programs: they can skip their entire larval stage [50]. This is presumed to provide an economic advantage in strong water currents as swimming larvae are not required for dispersal [50]. Clearly this alteration in embryology affects the pattern of cell lineage maturation [51, 52]. Sparse molecular
data have limited models of how different aspects of sponge development are integrated, however comparable basal models have been teased apart in relative detail.

**Molecular basis of basal body plan establishment: axis formation**

Planaria are basal model system in which molecular features establishing and maintaining its body plan are relatively well characterized. Planaria are flatworms with dozens of cell types that can regenerate adults from small body fragments of ~10 000 or more cells [53]. This ability is proposed to be a derivative of their asexual mode of reproduction [54, 55]. Indeed, withdrawal of the endocrine signals required to maintain regenerative capacity occurs coincident with the single episode of sexual reproduction that occurs in marine annelids [56]. Regeneration and asexual reproduction both require plasticity and coordination of maturation in sequences that do not occur during stereotyped progression from a zygote: outgrowth from ‘wound’ sites and reorganization of existing tissues called morphallaxis [53]. Morphallaxis is a cornerstone of the planaria plasticity program that also occurs in response to nutrient deprivation to downsize the organism [53].

Planaria use a constitutive gene-expression program of morphogen gradients that define the major axes and regional identities along them to achieve a consistent body plan [53]. Morphogen gradients endure lesions and asexual fission events, and provide a framework for regeneration and propagation of the body plan. The planaria SC population of neoblasts continuously replace differentiated cells through homeostatic cell turnover, where identities are defined according to these constitutive morphogen gradients [53]. These fundamental aspects of the body plan are molecularly conserved. For example, the dorsal-ventral axis of animals is often defined by Bone Morphogenetic Protein (Bmp) [57], and the Hox genes define regional identity
along the anterior-posterior axis. Indeed, the Hox genes are constitutively expressed in vertebrate skin, possibly to ensure consistency of positional information as it is continually renewed [58].

Establishment of a consistent body plan: lineage specification

It is a leap from a homeostatic molecular framework coordinating a consistent body plan to achieving the same endpoint from a single cell. Cell division, position and differentiation must all be coordinated. Emerging data are indicating that coordination of feed-forward loops accounts for much of the complexity, consistency and diversity observed in nature. In practice, feed-forward loops consist of TFs and ncRNA that coordinate transcription modules: early events trigger later ones and also feed-back to silence earlier ones.

The transition from a constitutive body plan maintained by morphogen gradients during asexual fission to development of complex body plans from single cells required that modules triggered by environmental signals during patterning of asexual body plans be activated by feed-forward loops during sexual development. For example in sponges, silicate induces the molecular machinery required for producing spicules (skeleton) as well as additional sclerocytes; this same program is activated genetically during development from a zygote [59]. Similarly, initiation of segment and tissue formation requires the presence of axes, but they are not propagated continuously in sexual reproduction. Various means of axes re-establishment are employed sexual reproduction. Symmetry is broken by environmental cues in sponges [42], asymmetric oocyte position in developing cysts and the ovary of *D. melanogaster* [60], sperm entry site in *C. elegans* [61] and possibly the mechanical pressure and spatial constraint exerted by ellipsoidal zona pellucidas on murine blastocysts that are restricted in shape by their outer trophoderm epithelium [62, 63]. Feed-forward programs then proceed into more specialized modules such as segment and tissue formation along the axes.
Feed-forward loops act at multiple levels of organization during development to coordinate cell position and fate. Using the seminal \textit{D. melanogaster} model [64], a feed-forward program established by maternal deposit proteins in the oocyte coordinates a series of anterior/posterior (A/P) defined segments consisting of thousands of cells. Reciprocal gradients of maternal deposit Nanos and Bicoid establish the A/P axis and initiate a feed-forward cascade: gap genes, followed by Hox and pair-rule genes, which in turn activate segment polarity genes [65]. Comparable feed-forward loops that consistently fate cells in appropriate positions are common in developmental programs.

\textit{C. elegans} have helped elucidate how molecular gradients are translated to arrays of highly specialized cell types. There are two underlying concepts. The first is that specific modules or programs are activated by combinations of TFs, such that a particular combination of TFs will activate a specific set of targets that define either a subsequent intermediate state or terminal differentiation program. The second is that cells undergo a series of binary ‘decisions’ while progressing through these intermediate states that culminates in a terminal differentiation program. By way of example, endoderm precursors in \textit{C. elegans} express a different combination of TFs through each division [66]. Binary decisions in this or comparable lineages in \textit{C. elegans} often use the same mechanism: asymmetric segregation of Wingless-related (Wnt)/Cadherin associated protein, $\beta$ ($\beta$-catenin) signalling components. A high SYS-1/POP-1 ratio in one daughter activates transcription, while a low SYS-1/POP-1 ratio in the other represses transcription [67-69]. This mode of achieving binary transcriptional output is combined with distinct TFs in many cell lineages during embryonic development, where these distinct TFs lend specificity in the set of genomic targets and ultimately alternative cell fates to the system [70]. The SYS-1/POP-1 system is not universal in establishment of alternative lineages \textit{C. elegans}, but
the same mode of asymmetric segregation to achieve alternate transcriptional programs and downstream fates is commonplace: the PAR (partitioning defective) system is used during the generation of blastomeres AB and P1 during the first cleavage of development [71]; Notch signalling in the subsequent division of AB [72] and asymmetric segregation of Ham-1 during divisions where one daughter undergoes apoptosis are alternate examples. Equivalent examples are evident in sea urchins [73], ascidians [74], and the nervous system (NS) of flies [75, 76] or vertebrates [77-80], suggesting the generality of this mode of generating cellular diversity.

‘Bottom-up’ approaches have established how intermediate and terminally differentiated cell types are defined molecularly. Continuing with *C. elegans* as a model, cell function or type can be defined as a combination of gene-batteries [70]. Each gene-battery is a functional unit, potentially comprised of multiple regulons, where each regulon consists of genes that are co-ordinately activated. For example, the N1 motif is present in the regulatory region of many ‘pan-neural’ genes [81], but disruption of other separate motifs dictating neuronal subtype will not interfere with pan-neural properties [73, 77, 82]. Transcriptome analysis suggests how function of neuronal subtypes is defined: a single gustatory neuron differentially expresses 68 TFs compared to a related sensory neuron, and the different regulons they activate the two subtypes [67]. Notably, targets are not exclusive to a regulon but activated by separate TFs such that targets are differentially combined according to the activators present. Repression of transcription plays a minor role, if any, in defining cell states in *C. elegans* [70]. Thus the functional properties of highly specialized or derived cell-types are the aggregate output of many activated regulons.

Cell states are stabilized by auto-regulatory activity and chromatin modification. Even after differentiation, cells will lose their terminally differentiated state after loss of terminal
selectors [67, 73]. Such identity defining TFs often maintain their own expression within a larger regulon. While use of auto-regulatory loops to maintain cell states is still common in vertebrates, chromatin modification provide an additional layer of stability. Histone modifications and DNA methylation that dictate access of TFs to regulatory elements are established during cell specification and inherited through mitosis to stabilize terminal cell states. Genomic chromatin maps have refined this model to suggest that lineage-specific loci are maintained as bivalent chromatin in pluripotent cells: a mixture of activating and repressing histone variants and modifications, before resolution during lineage differentiation to either activate or silent chromatin as appropriate [83]. Based on reprogramming studies where ‘terminal states’ are altered, models have been adjusted to suggest that progenitor cells represent metastable states and terminally differentiated cells represent stable TF networks that can be pushed to alternate lineages when the buffering capacity of a given TF network is exceeded intrinsically or extrinsically [80]. Notably in this view, multiple lineage transitions –e.g. alternatives to the developmental program- are possible [80]. These cellular states are endpoints for feed-forward loops, and provide stability for cells and the overall body plan.

Establishment of a consistent body plan: spatial coordination of lineages

The origin and utility of intercellular coordination during body plan establishment and lineage specification are illustrated in C. intestinalis, an ascidian that is slightly more complex than a sponge. In C. intestinalis, feed-forward activation begins with the maternal determinants β-catenin, Gata-binding (Gata) 4/5/6 and Macho1 that activate the majority of zygotic expression in the vegetal, animal and B5.1 regions of the embryo respectively. Six regions are established in the 16-cell embryo, with two modules (primary muscle and early epidermis) located furthest from signalling centers forming autonomously [49]. By the gastrula stage, expression of 53 of C.
intestinalis’ 600 TFs create a combinatorial code that is sufficient to specify all lineages [49]. However, the mode by which this combinatorial code is achieved does not rely as heavily on intrinsic feed-forward programs as C. elegans.

Coordination between modules during development is not an inherent requirement to achieve a consistent body plan. It is conceivable that the invariant lineages in ascidian develop from consistent segregation of maternal determinants across stereotyped cleavage planes; in fact, this hypothesis was the first proposal to explain the invariant lineages of ascidians [84, 85]. However the vast majority of cell fate decisions in ascidians are non-autonomous: stereotyped cleavage planes position competent cells to respond to paracrine signals from inducers [49]. Only 10 cells of the first 300 in ascidian embryogenesis form autonomously [49]. The remainder are induced by short range paracrine signals which confers positional constraint on development and may explain the degree to which development of distantly related ascidians exhibit conserved cleavage and spatial orientation of division events [49]. Indeed there appears to be greater conservation of spatial aspects of development in ascidians than in the molecular pathways [49], where paracrine and position-dependent cell-contact based signalling combine with intrinsic determinants to dictate cell fate. For example, all vegetal cells in C. intestinalis produce Fibroblast growth factor (FGF) and all animal cells contact a vegetal cell, however cell fate outcomes amongst animal cells are not uniform. Bipotential neural/epidermal cells are neurally induced if they reach a threshold of cell-cell contact with FGF releasing inducers [49]. Greater conservation of spatial than molecular programs holds when comparing ascidians to vertebrates which also have similar fate maps from the onset of gastrulation [49, 86].
Evolution of differing body plans

Body-plan organization can be categorized in tiered modules, where organizers, axes and segments provide the general framework. Alteration in the timing, distribution, levels and components of developmental modules are drivers of body plan diversity. Basal sponges have a single apical-basal axis, a ‘functional homologue’ of the vertebrate organizer called the oscule, and segmentation controlled by non-clustered paired-class genes (Pax-2/5/8) and LIM/homeodomain genes [39]. Two additional axes and control of segmentation by clustered homeotic genes have evolved from this basal state [39]. The significance of this latter shift is illustrated in classic experiments where Hox gene disruption led to homeotic transformations, e.g. conversion of segment identities [87]. Implicit in this observation is that segment development prior to Hox gene expression is equivalent such that different segments are still competent to enact alternative Hox-initiated programs. This is also evident ontogenically, as maggot segments do not exhibit diversity comparable to adults [88]. Characterization of homeotic mutants revealed that the Hox cluster is co-linear in that the genes appear in the same order in the genome that they are expressed temporally and spatially (A->P) in the organism [89]. Feed-forward expression occurs from each cluster where Hox genes define relative segment positions by repressing anterior structures [89]. Thus programs default to anterior appendages in the absence of Hox gene expression [88]. Genes are sequentially activated 3’ to 5,’ with each repressing a progressively more posterior organism segment and the Hox gene preceding it in expression [90]. As described by Ed Lewis before any Hox genes were cloned, “an antero-posterior gradient in repressor concentration along the embryo and a proximo-distal gradient along the chromosome in the affinities for repressor of each gene's cis-regulatory element [87].” The ancestral cluster is believed to have 10 linked genes producing a finite number of segments.
[88]. Competence for the structure produced in each segment is defined by distinct feed-forward cascades activated by each Hox gene that differ considerably between species [89]. Cooption of these segments and alteration in the timing and domains of their expression has contributed to the extensive diversification of body plans in the animal kingdom [89, 91]. For example, insect wings are believed to have evolved from the dorsal branches of an ancestrally branched arthropod limb [88].

Setting the balance between mosaic and regulative modes of development

The combination of feed-forward activation and coordination between lineages as modes by which cell fates are specified during multi-lineage differentiation resemble the descriptors used to categorize development: mosaic and regulative. For the purpose of this introduction, they are equivalent in distinguishing intrinsically and extrinsically driven processes. Asymmetric segregation of intrinsic determinants drive cell fate outcomes in mosaic modes of development, while cell fates are determined by non-autonomous signals in regulative development. Phylogeny suggests that regulative development preceded mosaics in evolution [50], exemplified by the regulative sponge at the base of the animal phylogenetic tree.

That mosaic development evolved in spite of the considerable benefits of developmental flexibility and environmental responsiveness offered by regulative approaches suggests considerable benefit. Comparison of three related worms offers insight into what this utility might be. C. elegans, A. nanus and E. brevis arose from a common ancestor, to which E. brevis is believed to be most similar. All three have nearly identical adult body plans, but differ considerably in their early developmental programs. Development of E. brevis is largely regulative with variant cell lineages, some capacity for regeneration, and no early apparent germline [50, 92, 93]. Conversely, C. elegans is largely mosaic and A. nanus an intermediate...
One notable difference in the developmental programs is that *C. elegans* develops in ~1/5 the time of *A. nanus*, despite eggs that are approximately the same size and larvae hatching with approximately the same number of cells [50]. Illustrations of selective pressure often relate to efficiencies in adult life, however reducing the duration of embryogenesis facilitates expansion of populations, particularly relative to another with a comparable body plan. This efficiency comes at the expense of regulative properties: in *A. nanus* development is relatively slow and inhibitory signals are required to prevent ectopic differentiation [94], but cell losses can be compensated from alternate lineages during development. Comparable plasticity is absent in *C. elegans*, as lost cells are not replaced [95]. Condensing development might involve increased maternal investment, but in this comparison is believed to coincide with colonization of terrestrial habitats by *C. elegans* where resources were plentiful, but variable in their availability compared to a stable marine biotope. This example illustrates comparable influence of selective pressures in moulding development to the more familiar examples of moulding adult body plans.

Curiously, comparable reduction in the duration of each stage of development did not occur between these three species: the differences between developmental programs are most pronounced prior to gastrulation. This observation indicates that not all stages of development are equally facile in their responsiveness to selective pressure. Constraint on developmental stages generally occurs when the level of interactions between developing modules is highest. If the stage is sufficiently constrained that it corresponds to a period of embryogenesis when different species within a clade are most visually similar, it is referred to as a ‘phylotypic stage’ [52]. The segmented germ band stage that occurs after gastrulation is considered the phylotypic stage in arthropods [52, 96], whereas the pharyngula of chordates is phylotypic. Molecular data have provided insight into the basis of these observations: variation in expression levels and
sequences of genes expressed in phylotypic stages are reduced relative to gene expression during other embryonic stages and sequences of genes expressed during these other timepoints [97, 98]. The prior example of ascidian development illustrates spatial constraint in developmental: in spite of highly diverged molecular signals, the proximity of inducing/responding cellular pairs is conserved between divergent species [49]. The inference is that alterations are less likely to improve dependent processes, e.g. those requiring coordination between modules, than autonomous processes. Perhaps it is counterintuitive then that regulative development, illustrated by the ancestral *E. brevis* in the previous example, is intrinsically more amenable to cooption than mosaic development. Coordination by extrinsic signals certainly does occur in regulative development, but compensation for deviations in the developmental program is an inherent feature of regulative development, whilst the precision of mosaic development—in terms of coordinating timing and position of separate intrinsically driven lineages—is less amenable to compensation, and hence cooption.

While organism development preceding gastrulation is not highly constrained, certain overt similarities are evident in terms of rapid cellularization and epithelial formation [99]. Development prior to these common features varies considerably in terms of reliance on zygotic transcription, the mode and pattern of cell division, as well as coordination of cellular expansion with differentiation.

In contrast to *A. nanus* which requires transcription for cleavage divisions, *C. elegans* produces 100 cells in the absence of zygotic transcription [94]. In an even more extreme comparison, *X. laevis* embryos undergo 10 rounds of division before zygotic transcription [100]. At the outset of development, asymmetric divisions in *C. elegans* embryogenesis, symmetric cleavage divisions in mammalian embryogenesis, and absence of cytokinesis early in
embryogenesis of many arthropods, such as *D. melanogaster*, illustrate the variety in modes of division. *D. melanogaster* embryos undergo repeated genomic replication in a syncytium: 13 synchronous cycles in ~2 hrs to produce ~5000 nuclei, before cellularization and formation of an epithelium [99, 101, 102].

The extent of expansion of undifferentiated cells and lineage specification at the outset of embryogenesis is similarly diverse. In 16-cell ascidians, one of the 6 defined domains is already restricted to the epidermal lineage [103]. By the onset of ascidian gastrulation there are ~100 cells in the embryo, and 38/44 endoderm/mesoderm precursors are fate restricted [49]. By comparison murine blastocysts of 100 cells are primarily composed of extraembryonic trophectoderm (TE), but of the fraction allocated to form the embryo none are thought to be lineage restricted. The pluripotent ICM comprises ~20-30 of the cells in 100-cell blastocyst [104], before further expansion as epiblast and gastrulation. While the murine embryo does make a substantial contribution to extra-embryonic tissues, the epiblast which will generate the embryo remains pluripotent.

**Pluripotency in mice**

The molecular network responsible for pluripotency is the starting point from which all terminal somatic differentiation programs are initiated. The static state of diapause, entered into 3-4 days post coitum (dpc) by murine pre-implantation embryos if their mother is malnourished [105-108], has facilitated characterization of pluripotency. This static state and responsiveness to extrinsic signals during cell fate specification illustrate the regulative nature of murine development. Outgrowths of the pluripotent inner cell mass (ICM) on a layer of fibroblasts in media containing fetal bovine serum were initially used to capture self-renewing pluripotent embryonic stem (ES) cells, an *in vitro* derivative of the ICM [109]. Later studies revealed that
pluripotency depended on Leukemia Inhibitory factor (LIF) secreted by the fibroblasts [110] and Bmp4 from serum [111]. Since then, it has been established that extrinsic factors are only required to prevent differentiation induced by paracrine factors released by ES cells [112]. Indeed, if extrinsic factors are negated, the ground state of pluripotency is believed to be stable and intrinsically maintained [113]. Extensive characterization has revealed key features of how a ‘distinct higher-order global chromatin structure,’ and intrinsic pluripotency network maintain the undifferentiated state [114]. The network consists of auto-regulatory TFs that maintain stability of the pluripotent state. The core components --Oct4, Sry-box containing gene 2 (Sox2) and Nanog homeobox (Nanog)-- each regulate their own expression as well as that of each other in an arrangement referred to as a fully connected triad [115-118]. The network is highly buffered by redundant components, ncRNAs as well as overlapping regulons [115]. Many components of the network also interact physically [119-122], suggesting either cooperative or regulatory relationships.

Components of the pluripotency network have been defined by a variety of approaches, with requirement of a factor for pluripotency being the most straightforward. Redundancy is one caveat to this. For example, it was not revealed that Kruppel-like factors (Klf) were essential for pluripotency until klf2, klf3 and klf4 were all removed simultaneously [123]. Sox2 illustrates a different scenario. Its categorization as an essential factor on the basis that its disruption leads to loss of pluripotency fits the general case. However of the thousands of putative genomic targets Sox2 has in ES cells, its only function relevant to maintaining pluripotency is proposed to be maintenance of Pou5f1 expression (Pou5f1 encodes Oct4) [124]. Sox2 is often studied in conjunction with Oct4 because their functions are believed to be intimately linked. They are cofactors --derived from an ancestral HMG domain protein and homeodomain protein
respectively— that are believed to have co-evolved interdependent roles in regulating transcription [125]. Recurrence of adjacent consensus binding sites of Sox2 and Pou5f1 throughout the genome suggested overlapping regulons and related function [126]. Physical co-occupancy of the two at many genomic targets supported the suggestion that Sox2 and Oct4 bind cooperatively to activate or repress transcription that was initially made through study of the Fgf4 locus [127]. Furthermore, functional data from other SC populations where Oct4 is absent indicated Oct4-independent requirements of Sox2 in maintenance of neural stem cells (NSCs) [128], retinal progenitor [129], and osteoblasts [130]. Extensive data support cooperative roles of Oct4 and Sox2, as well as an Oct4-independent role of Sox2 in SC maintenance, but functional data conflict and suggest that the sole function of Sox2 in pluripotency is maintenance of Pou5f1.

The collective function of the pluripotency network is generally modelled as buffering against differentiation by limiting target activation of lineage defining factors [113, 131]. However emerging data suggest that this model may not capture the complete picture. ncRNAs (ncRNAs), some of which maintain the rapid cell-cycle of ES cells, are critical to maintenance of pluripotency [132, 133]; heterogeneity in expression amongst pluripotent populations—such as oscillations in Nanog—suggest transition through various states [134]; and the alternative suggestion that rather than limiting activation by lineage specifiers, the pluripotency factors are themselves lineage specifiers balanced by mutual repression [135], are considerations that are not fully accommodated or resolved by existing data. Similarly, molecular heterogeneity in pluripotent populations [136], and mosaic differentiation beg the question of whether functional heterogeneity exists in these populations [137].
A combinatorial code of pluripotency factors is proposed to dictate target activation versus repression, akin to the combinatorial code in lineage specification of *C. elegans* described earlier, though no code has been demonstrated that strictly correlates with either. At present the strongest correlations are between the number of TFs at a given locus and activation, although these are not strict, even when using several factors [126, 138]. Oct4 illustrates this unknown. It is a TF that both activates and represses transcription [139]. It required for pluripotency of the ICM and ES, which both differentiate into TE in its absence. It is sufficient to induce pluripotency in NSCs on its own [140], and in combination with other TFs induces pluripotency in somatic populations [141]. Yet it is still not clear how Oct4 contributes to pluripotency, e.g. no rescue has been performed with downstream factors in the way that Oct4 production rescues Sox2 depletion to confirm its essential contribution. The general tact to answering this question has been to define the contribution of factors to pluripotency by relationships of essential factors: physical associations between TFs required for pluripotency or common genomic targets. Such network analysis often encompasses both direct and indirect associations. However these networks are based on proliferative ES cells and the only instance when pluripotent cells are maintained in a consistent state for more than a couple of days during development is when division of the ICM is arrested in diapause. Ordinarily, the ICM is a transient structure. Oct4 is present for several days of development after the last pluripotent cells are believed to be present in the embryo [142], suggesting it is insufficient for pluripotency in these contexts. Since Oct4 persists after pluripotent populations and its depletion appears coordinated with differentiation, Oct4 has also been suggested to play a role in cell fate outcomes [143, 144].

In this thesis I ask what role Oct4 has in somatic cells after the onset of gastrulation. The *Pou5f1* regulatory element is amongst the most complex in the genome in terms of the
number of factors which bind it, so one possibility is that it is activated by an alternate portion of the regulatory element to achieve multiple context-defined functions. In this scenario, Oct4 would mediate different regulons in different contexts. Alternatively, it may have the same function and regulon outside of the pluripotency network, but contribute to a different cell state via combination with different regulons that are not part of the pluripotency network.

Tradeoffs in selective pressures defining cellular lineages and body plans

The morphogenetic movements of gastrulation produce multilayered embryos with specified lineages. Gastrulated chordates develop into neurula and then pharyngula –the chordate phylotypic stage– before species-specific programs diverge according to the selective pressures that shaped their respective modes of development and body plans. The array of mature ‘adults’ that result from this divergence is enormous, with an estimated 1.5 million species of animals alone [145]. Consistent optimization, or evolution of progressively more specialized features within species does not seem compatible with compromise, however intrinsic and extrinsic factors influencing survival and reproduction mould body plans through tradeoffs: an adaptation that is detrimental in one context will be retained if the benefit it provides enhances overall survival or reproductive success. Tradeoffs incorporate a range of factors, from genome integrity at the cellular level to optimized body plans.

The influence of sexual reproduction on body plans

Tradeoffs are prevalent in nature. Inter-species comparisons might suggest that specialization is incompatible with tradeoffs: camouflage is prevalent, as are vibrant colour schemes that distinguish species from their environment. These suggest dominant influence of either predator evasion and prey capture or mate-attraction on color schemes. However a tradeoff
in bats illustrates the norm of how competing selective pressures are balanced. Flight constrains resource allocation in the bat body plan, leading to a tradeoff, rather than independent specialization of single tissues in response to selective pressure. Competition between sperm from multiple mates occurs following promiscuous behaviour offers one illustration. In species where female bats are promiscuous, a trade-off occurs between the two most metabolically active tissues: male brains shrink and their testicles grow to improve odds of winning the sperm competition [146]. A similar situation was summarized by Jon Lien who noted that the highly promiscuous whale “known as the black right whale has four kilos of brains and 1000 kilos of testicles. If it thinks at all, we know what it is thinking about” [147]. In bat species where females are monogamous, the male body plan shifts to larger brains and smaller testes, where larger brains are associated with increased survival [148].

Despite negative consequences in certain regards, sexually dimorphic traits are not maladaptive overall. Vibrant color schemes associated with mate attraction increase susceptibility to predation, but the associated increase in reproductive success balances this disadvantage. But why are traits that are disadvantageous by some criteria used at all in mate selection?

Sexual dimorphism encourages the benefits conferred by sexual reproduction. Sexual modes of reproduction are believed to provide utility in enhancing the rate of trait adoption (e.g. trait-fixation occurs sequentially in asexual populations, but can occur in parallel in large sexually reproducing populations [149]), facilitate selection of genetic synergies by combining different variants each generation and generating diversity amongst littermates through meiotic recombination to resist parasitism [150]. This latter point is particularly relevant. Since parasites often have faster generation times than their hosts, genetic diversity amongst offspring limits the
spread of pathogens. By extension, if speciation is driven by sexual dimorphism (which is debated [151, 152]), this could also confer resistance to pathogens through isolation. The resistance against parasites is far from complete in sexually reproducing species, as genomic parasites (transposable elements) thrive in this context. While transposition events are generally deleterious and will limit expansion of clonal asexual populations, deleterious transposition events are tolerated in sexual reproducing offspring because they effect a minority of the offspring [153]. Overall, insurance of genetic diversity in offspring is a sufficient offset for maladaptive aspects of sexual reproduction.

The influence of intrinsic cellular constraints on body plans

Do intrinsic limitations such as genome integrity and cellular senescence constrain complex body plans? Consistent development of body plans requires faithful replication of the genome through each cell cycle. This is apparent in the absence of RNA-based multicellular organisms, and expansion of repair and checkpoint mechanisms in complex organisms [154]. Alterations in replication fidelity have been proposed to optimize the species-specific rate of evolution [155], however given that mutations are overwhelmingly deleterious a model suggesting that higher mutation rates are selected is tenuous. Genetic variation is needed for selective pressure to act on and retrieve the occasional beneficial mutation, so this advantage coupled with the energetic cost associated with fidelity do establish a lower limit on the error rate during replication [156]. Indeed, in a surprising observation, a near consistent rate of mutation per genome in single-cell microbes whose genomes range over nearly four orders of magnitude provides strong evidence that an equilibrium setting the mutation rate per genome exists [156]. That said, the general trend towards greater fidelity per base with organism and genome size supports selective pressure for fidelity as organism complexity increases.
Constraint by cellular senescence predicts that expansionary divisions should occur uniformly amongst progenitors to avoid the Hayflick limit of ~40-60 divisions by any particular cell [157]. As described earlier, the uniform symmetric expansion predicted by this constraint is a hallmark of early development, however ‘bottlenecks’ in cellular lineages occur roughly coincident with cell differentiation. For example, considerable apoptosis occurs in the *D. melanogaster* neuronal population following pupariation and eclosion of adults [158], and unsynapsed neurons apoptose in response to insufficient NGF signalling in mammals [159]. These are inconsistent with constraint by cellular senescence.

Rather than dominant influences of cellular constraints, development appears to abide by the same general principle that body plans do in that programs proceed according to utility. Instead of uniform symmetric expansion followed by synchronous differentiation, allocation, and in some cases differentiation, occur long before the end of embryogenesis. For example, in murine development, cardiomyocytes are specified and beat to meet circulatory requirements before posterior limb buds even form. Furthermore, rather than being a detriment, expansion driven by a subset of asymmetrically dividing cells suggests this mode is an asset and not a liability in developmental programs.

Asymmetric division facilitates genuine self-renewal in a subset of cells, ensuring faithful development of complex organisms. Retention of the template genomic strand via asymmetric division [160], cellular quiescence [161] (although cellular senescence is not strictly correlated with replication [162]) and genome-surveillance [163] are modes by which genome integrity can be maintained. In addition, asymmetric partitioning of macromolecules damaged by oxidative stress to one daughter cell of a division facilitates bona fide self-renewal by providing the other daughter with undamaged cytoplasmic components [164]. A final consideration of quality
control within the each organism is cell competition. Cell competition is proposed to enable the ‘fittest’ progenitors to make the largest contribution to their respective tissues [165, 166]. All of these modes of ensuring cellular integrity are characteristic of SCs. SCs contribute to development and enable plasticity throughout the lifetime of an animal. They underlie repair, regeneration, learning and memory as well as adaptive immunity. Adult SCs are, in effect, an adaptation towards adaptability.

These principles are illustrated in the NS and germ line.

Evolution of the Nervous System

Evolution of a highly derived cortex underlies the age old question of what it means to be human. An early theory on the origin of the NS released by George Parker in 1919 persisted for the first half of the 20th century. It suggested that neural connections evolved to coordinate ‘independent effectors’ in response to stimuli (for example contractile cells in sponges where there is no NS) before derivation to process additional sensory information and coordinate additional behaviour [167, 168]. Neurons were proposed to arise from superficial epithelial cells [167, 168]; intercellular communication via junctions would already have been present, and apical/basal polarity encompassing localized phagocytosis and intracellular trafficking would be further derived to dendrites and axons. The first neurons were proposed to excite adjacent muscle cells, and also extend processes to propagate excitation through nerve-nets as is seen in cnidarians [167, 168]. The cnidarian NS is diffuse, lacking directionality, without defined axons or dendrites, and combining sensory and motor functions within cells [167].

There have been several refinements to this model. The first being that enhanced coupling between stimuli and behavioural responses may not have been the initial selective pressure prompting adoption of a NS. Basal syphomedusae and hydra possess neural pacemakers
prompting the hypothesis that the earliest neural networks coordinated rhythmic contractions at intrinsically dictated rates as opposed to being responsive to external stimuli [169]. A second being that neural networks may not have evolved to connect existing pairs of neurons and contractile cells, but been a refinement in connectivity of syncytial epithelium [167]. Indeed, action potential conducting epithelia are present in hydromedusae and siphonophores [167], so neural nets may have evolved by reducing connectivity within these syncytia to refine signalling [167]. A third is that the defined neuronal morphology of elongated processes evolved in secretory cells, with concentrated receptors at one pole and secretory apparatus at the other, before cooption to electrical transmission [167, 170]. Finally, the NS was recently proposed to have ‘centralized’ several times independently in evolution, and separate types of neurons to have evolved independently [171]. The latter claim is based on the lack of pan-neural genes in all phyla and neurons originating from different germ layers in different phyla [171].

Consistent utilization of genetic tools in the NS of diverse phyla are used to support fixation of the neural program in a common ancestor in the former theory and explained by convergent evolution in the latter [171]. Neural specification by inhibition of the mesoderm ventralizing signal BMP is conserved from vertebrates to arthropods [172, 173]. In mice, the BMP inhibitors Noggin and Chordin are released from the organizer [174], while Follistatin is also used in some other species [175]. Conserved TFs indicate specification of neuroepithelium (NE). In D. melanogaster these include genes dichaete, achaete, scute and lethal of scute [176-179], which are recognizable in similar roles in mice: dichaete is related to Sox-1, 2, 3 which are essential for development of the murine NS [180-182], while Ascl1 (achaete-scute complex homolog 1) is required early in select murine neuronal differentiation programs [183].
Contribution of neuroepithelial organization to cortical evolution

Species-specific differences in the NE contribute to cortical differences. Neuroblasts (Nb) are specified by waves of proneural gene expression in the ventral NE of *D. melanogaster*. These waves are defined by an intrinsic temporal feed-forward program that produces distinct Nb before they lose their intercellular junctions and delaminate to the interior [184, 185]. Subsequently Nb self-renew via asymmetric divisions basal to the apical surface of NE and produce ganglion mother cells that divide once more to produce neurons and/or glia [186]. NE and Nb both have apical/basal polarity, but spindle pole alignment parallel to the apical surface leads to symmetric divisions in NE and perpendicular alignment to asymmetric Nb divisions [184]. Different signalling cues present during each wave of Nb division may contribute to neuronal diversity [187]. Neuron number varies between species of insects according to the number of Nb divisions, where the initial number of Nbs is similar between different insect species [186]. As in insects, the number of Nb divisions varies considerably in crustaceans [186, 188, 189]. However a difference arose in ancestrally higher crustaceans in that Nbs do not delaminate to the interior. Instead, they divide perpendicular to the neuroepithelial surface and produce ganglion mother cells in columns that are progressively pushed towards the interior by each subsequent Nb division [186].

Two other apical progenitors accompany the NE in rodents: radial glia and short neural precursors [184]. Each of these cells has hallmarks of epithelial morphology (apicobasal polarity, apical junctional complexes and radial bipolar morphology), undergoes interkinetic nuclear migration (apicobasal movement of the nucleus during G2 with mitosis at the apical surface [190]) and an apical process that contacts the ventricular lumen [184]. Rodents also have basal neural progenitors. They are referred to as intermediate, non-surface or subventricular (SVZ)
progenitors, found almost exclusively in the telencephalon and primarily divide symmetrically [191, 192]. Symmetric divisions generally result in two neurons, but in 10% of cases yield two SVZ progenitors [192, 193]. The absence of apicobasal polarity may explain symmetric division of SVZ progenitors.

Cortical expansion can occur by radial or lateral expansion. The cortex is built by progressive layering of neurons from the inside out, so radial expansion in evolution might consist of additional layers or their thickening. In lateral expansion, additional neurons are generated adjacent to one another. Unlike insects where an increase in neuron number is associated with additional rounds of division by each Nb, mammalian cortices have expanded laterally [184, 194, 195]; this is evident from an increase in the size of the external cortical surface relative to the ventricular membrane [196], and suggests an increase in the number of Nb via early symmetric expansion [184]. Lateral expansion in rodents is accomplished in part by evolution of a pseudostratified NE. Unlike D. melanogaster, for example, where symmetric NE divisions and asymmetric Nb divisions occur in different layers, the pseudostratified NE of rodents facilitates expansionary symmetric divisions and asymmetric divisions in the same junction-connected layer [197].

Interspecies comparisons of vertebrates and neuroanatomy indicate that increased brain size was accompanied by an increase in the external pial layer relative to the ventricular surface as well as by increased density of progenitors on the apical surface [184]. Increased density at the apical surface is accommodated by elongation of apical progenitors such that each extends a primary cilium to the neural lumen [198, 199]. As such nuclei are stacked – basally located during interphase and apically located during mitosis [190]. Self-renewal requires that both daughter cells inherit a portion of the apical membrane [200-205], so elongation and the
subsequent decrease in luminal surface contact by each cell necessitated tighter coordination of division planes [184]. Apical mitoses may be constrained by requirement of the luminal centrosome from the primary cilia for mitosis [184, 206]. Freedom of mitosis at the apical surface from steric constraint by cell elongation and stacking of nuclei increased the NE’s proliferative potential [184].

Division mode in the SVZ and its relation to cortical expansion

Elongation of pseudostratified NE may have altered the default mode of division from symmetric in basal species where cells comprising the NE have large apical domains, to asymmetric. Indeed, in *D. melanogaster* the transition from symmetric to asymmetric divisions is an active process requiring Inscuteable (Insc) [207, 208]. Insc promotes horizontal cleavage planes in murine epithelium suggesting its role in mediation of spindle orientation is conserved [184, 209]. The predominantly asymmetric vertical cleavage planes in the mammalian cerebral cortex may relate to the absence of Insc. Insc is absent during the transition from symmetric to asymmetric divisions [191, 205, 210], and forced expression increases the number of symmetric divisions [204]. This shift in mode of division can occur independent of altered spindle orientation however, suggesting that alternate pathways may be causal [184]. Nonetheless, if evolution of pseudostratified NE made asymmetric division the default mode, tighter regulation of spindle orientation would likely be required to promote symmetric divisions [184].

Adult neurogenesis

The majority of neurogenesis in the mammalian cortex ceases shortly after birth [211], however the subependyma (SE; adult descendent of the embryonic germinal zone) and subgranular zone (SGZ) of the dentate gyrus are notable in that they harbour neurogenic populations in adult mammals. New neurons, or neuroblasts, produced in the mammalian SE
migrate along the rostral migratory stream to the olfactory bulb where they differentiate into granular and periglomerular interneurons [212]. Neurogenesis in the dentate gyrus generates primarily local glutamatergic excitatory dentate granule cells [213, 214].

Cnidarians may be the most basal organism to utilize neuronal replacement [145, 215]. In crickets, 20% of the neuronal Kenyon cells in the mushroom body (an integrative center for multimodal inputs) are born in adults and contribute to its remodelling [216]. In decapods crustaceans, neurogenesis is continuous through lifespan in the olfactory pathway. Teleosts exhibit neurogenesis throughout the CNS [145, 217]. Continuous growth of both decapods crustaceans and teleosts as adults may explain their extensive adult neurogenesis. Finally, neurogenesis occurs in all vertebrates examined to date [145]. While neurogenesis in the periventricular zone (PVZ) is conserved amongst vertebrates [145], the distribution of mammalian neurogenesis represents a restriction in both the number of neurons born in adults and the breadth of neurogenic regions [145].

Neural stem cells (NSCs) in the mammalian SE produce interneurons that relay sensory information from the field of olfactory neurons. The neurons in the olfactory field are heterogeneous for olfactory receptors and the interneurons produced during adult neurogenesis are thought to adjust sensitivity of these olfactory neurons according to usage [218]. Neurogenesis in the SGZ produces neurons from the hippocampus that enhance consolidation of spatial memories and mood [213, 214, 219].

The ontogeny of NSCs responsible for neurogenesis is only partially understood. NSCs are derived from radial glia [220], and share some markers such as Glial Fibrillary Acidic Protein (GFAP) and Glycoprotein CD133 [214]. However it is not known how prospective NSCs are sequestered from radial glia – induction within a subset that are broadly competent, or selection
based on access to a niche or other criteria. Fundamental questions pertaining to NSCs remain. The extent of their function and the molecular network maintaining their potency are unknown. Some components of the intrinsic NSC program are established, for example Sox2 and T-cell leukemia (TLX) are required for NSC maintenance and proliferation [221, 222], while Pax6 is required for maintenance and neuronal differentiation [223]. Extrinsically, FGF, Epidermal Growth Factor (EGF), Wnt, Sonic Hedgehog (Shh), BMP antagonists, Notch, LIF, Transforming growth factor (TGF)-α, cytokines and Brain-derived Neurotrophic Factor (BDNF) facilitate neuralization, define competence, promote proliferation, maintenance, survival and differentiation of NSCs and their progeny [228,229,230[214]]. Maturation within a lineage requires that these intrinsic and extrinsic signals be coordinated with other unknown signals. In this thesis, I ask how early phases of NSC lineage maturation prime extrinsic responsiveness for later sequential stages of maturation and coordination with other aspects of development.

Evolution of the germ line from somatic SC populations

Due to interrogation bias in the order germ lines of model organisms were studied, ‘preformation’ was erroneously presumed to be the ancestral form of germ line development. In preformation, the germ line is established by immediate sequestration from the soma following fertilization [224]. Maternally deposited germ-plasm is asymmetrically segregated into the presumptive germ line where it prevents transcriptional responses to somatic inductive signals. In C. elegans, D. melanogaster and X. laevis, transcriptional responses are inhibited by preventing RNA polymerase II activity [224]. Studies encompassing additional model organisms have revealed that an alternate mode of germ line development, ‘epigenesis,’ is ancestral. Epigenesis refers to induction of primordial germ cells (PGCs) from pluripotent cells. In basal
metazoans (sponges and cnidarians) as well as bilaterians (flatworms), gametes and somatic cells are produced from the somatic SCs [42] (cnidarians: [225-227]; planaria [228-230]). Sufficient potency for functional germ cells is present in somatic SCs of planaria, evidenced by regeneration of functional germ lines in fragments devoid of such structures, and by induction of germ cells from neoblasts in sexual hermaphrodites [4, 231]. Furthermore, conversion of asexual to sexual planaria is feasible experimentally (it is thought that planaria oscillate between modes of reproduction seasonally in the wild, but not demonstrated) [231, 232]. Germ cells in complex organisms are believed to have evolved from somatic SCs in basal organisms.

Markers of oocytes and gemmules in sponges offer insight into the origin of reproductive programs in complex organisms. A homologue of Polycomb repressive complex 2 (PRC2) member EED (Embryonic Ectoderm Development) marks gemmules, and a RTK homologous to the EGF intracellular domain marks oocytes [39]. PRC2 mutants in complex organisms are sterile, supporting retention of PRC2 requirement in derived germ lines: (C. elegans: [233] and D. melanogaster [234]). Similarly, the Piwi/Piwi-interacting RNAs (piRNAs) system required for germ line maintenance in many complex species maintains pluripotency of somatic SCs in planaria and may do the same in jellyfish, cnidaria and sponges: both Archeocyte and Choanocytes contain Piwi homologues [42]. Other genes present in both germ and somatic SCs of basal metazoans that appear restricted to roles in establishing or maintaining the germ line of complex metazoans include: Vasa, which encodes a DEAD-box RNA helicase (planaria: [235], hydra: [227]), and Nanos, a CCHC zinc finger RNA-binding protein expressed in germ line SCs (in planaria: [236], in hydra: [226]) [42]. Nanos and Vasa homologues in hydra are upregulated during differentiation of germ line SCs from basal levels of expression in somatic SCs [226], suggesting a molecular switch that may have facilitated the establishment of germ line SCs in
basal organisms. The presence of molecular machinery in somatic SCs of basal multicellular ancestors that has been restricted to germ line maintenance in complex organisms supports the evolution of gametes from somatic SCs.

As the sole cellular lineage contributing heritable information, the germ line experiences unique selective pressure. One major adaptation is the derived state of preformation, which is thought to have arisen independently several times to enable somatic lineage differentiation earlier in development thereby condensing the duration of embryogenesis [224]. There are several corollaries to this. If pluripotency is ancestral, it should be conserved. Repression of somatic differentiation signals in the germ line would facilitate early lineage differentiation in preforming species. And relief of constraint would occur in somatic differentiation programs shared with the germ line specification program in performing species. These hold true. Orthologues of the pluripotency factors Oct4 and Nanog are present in the animal cap of ancestral axolotl (epigenesis) [237]. Conversely, independence from the pluripotency network associated with use of germ-plasm may explain loss of Nanog from the teleost and X. laevis genomes (preformation) [224]. Oct4 does not fit this model as cleanly. In agreement with predictions, the axolotyl orthologue can partially compensate for loss of murine Oct4, while the D. rerio paralogue cannot. However, compensation by the X. laevis paralogue for murine Oct4 is not predicted. This does not conflict with this model, but suggests that the conserved function may not be pluripotency.

Relief of constraint on body plans and lineage specification programs correlates with adoption of germ-plasm. Evolution of germ-plasm precedes enhanced speciation which enables liberation of body plan patterning earlier in development by shielding the germ line from somatic inductive signals [224]. The function of genes is often conserved between lineage-specific
somatic programs of preformation and epigenesis-employing species, but the regulatory relationships are in performing species are not constrained to the same extent. For example, TGF-β and FGF signalling specify mesoderm —and the germ line in epigenesis-employing species. The ancestral state involves a cascade of single genes and is conserved from axylotyls to mammals. Conversely, while these components are used in X. laevis, the network consists of over twenty-five copies of Nodal and seven of Mix [238].

Selective pressures unique to the germ line continue after its specification. The differential contribution to reproduction made by the two sexes underlies another selective pressure. As described in Bateman’s principle [239], females are rate-limiting in reproduction as they generally contribute more to offspring than males. There are several corollaries to this. Since male, but not female, promiscuity correlates with offspring number, females engage in mate selection while males engage in promiscuity. Since there is greater variance in male than female reproductive success, males take risks associated with being at the ‘successful’ end of the broad spectrum of male reproduction frequency. This accounts for the elaborate plumage and intra-species confrontations that occur primarily between males. Sex-specific selective pressures are continuous from mating throughout child rearing.

Imprinting

Imprinting refers to parent-of-origin dependent inheritance of epigenetic features that lead to allelic-expression bias: either the mother or father’s allele is preferentially expressed since expression of the other allele has been repressed. The ‘imprint’ is the epigenetic mark that decreases expression in one of the two germ lines, often impacting a cluster of adjacent genes directly through cis-acting effects mediated by processive transcription or ncRNAs. Parent-of-
origin dependent traits are sexually dimorphic in that they ultimately reflect differential selective pressure on males and females in terms of the cost of reproduction.

The advantages afforded to physically large mammalian offspring and disadvantages afforded to mothers of excessively large offspring are competing interests that form the basis of the ‘conflict hypothesis’ [240, 241]. Large offspring have advantages in both survival and reproduction, so programming sperm to suppress growth suppressors --a gene expression program that plays out during embryogenesis-- offers their offspring a clear advantage. To a mammalian mother, which bears the energetic cost of this growth via nutrient exchange across the placenta or milk provision, excessive offspring growth might be taxing to the point that it impairs their survival during pregnancy, fitness for future pregnancies or places her survival at risk during delivery. These opposing pressures were postulated before the discovery of imprinted expression [241]. The first genomic screen for such traits, reciprocally mated mice carrying uniparental disomies, appeared to validate this hypothesis by revealing imprinting of several growth regulators [242]. The first set of ‘opposing’ phenotypes observed following inheritance of uniparental disomies from opposite parents concerned a Robertsonian translocation on chromosome 11 [242]. Embryos inheriting maternal segments of chromosome 11 were consistently smaller than embryos with standard diparental inheritance of the segment, whilst embryos inheriting paternal segments were consistently larger [242]. Uniparental inheritance of the distal region of chromosome 2 provided the second example of opposing phenotypes in that maternal duplication resulted in progeny with ‘long flat-sided bodies’ that were nearly inactive following birth, whilst paternal inheritance results in ‘short square bodies’ and hyperkinetic behaviour [242]. In both cases the phenotypes fit the ‘conflict hypothesis’ in that inheritance of two paternal alleles led to overgrowth. Continued research eventually demonstrated that several
imprinted genes were found in within a cluster in both cases. The cluster on chromosome 11 consists of $Ddc$, $Grb10$, $Cobl$, $Zrsr1$ and $Commd1$, while the cluster on chromosome 2 consists of $Gnas$, $Gnasxl$, $Nespas$ and $Nesp$. Imprinting of $Grb10$ is likely to be primarily responsible for the former phenotype [243], and imprinting of $Gnas$ for the later [244]. Phylogenetic studies of the canonical imprinted gene $Insulin growth factor 2 (Igf2)$ assessing when its imprinting became fixed in evolutionary terms seemingly clinched this model. $Igf2$ is imprinted in all viviparous mammals where resource allocation is relatively plastic, but not in monotremes or birds where egg size is fixed before fertilization [245].

There are issues with the central tenets of this hypothesis. For example, the conflict hypothesis suggests that genes should not be imprinted in monogamous species, but they are [246]. The ‘direction’ of parental bias is also opposite predictions in some cases, such as Prader-Willi syndrome where a paternal deletion causes over-eating and obesity [247], in contrast to the model’s prediction of growth reduction when function of a paternally expressed gene is disrupted. Since imprinted expression has been assayed with approaches dependent on visually apparent phenotypes (growth being one of the most obvious), unbiased approaches have revealed numerous loci that are imprinted, with no apparent role in growth. For example, approximately 20% of the known imprinted genes do not have growth phenotypes, but are implicated in neurological processes [248]. These do not all influence provision of milk as would be predicted by the conflict hypothesis [248]. Another example of imprinting that is thought to be unrelated to growth are a class that have undergone retrotransposition from the X to autosomal chromosomes [249]. Imprinting may serve the same purpose as X-inactivation in ensuring appropriate dosage of these genes [249]. Assays for imprinting have shown that imprinted loci vary considerably between species [250]. Thus testing one locus –albeit $Igf2$ which is conserved in viviparous
mammals—to demonstrate that selective pressure for imprinting is absent in monotremes and birds is inadequate since it assumes the selective pressure that prompted imprinting is growth regulation. Absence of Igf2 imprinting in egg-laying species could be used to support the position that flexible nutrient exchange across the placenta was the forum in which selective pressure for imprinting arose, although monotremes do have a short-lived placenta [248]. This position would be further supported by the high density of imprinting in the placenta [251]. Conversely, lactation is also flexible in monotremes as is incubation of eggs in birds that do not equally share the responsibility, so the conflict hypothesis would predict that imprinting should have also arisen in these contexts. If imprinting did arise in these species, it would occur in different tissues and at different genomic loci. A systematic and unbiased approach is needed in non-viviparous species to rule out the possibility that imprinting was selected by another pressure and has been coopted differently between animal classes.

Imprinted loci vary considerably between species [251], so unless the catalogue of genes which are imprinted in each species is calibrated to reflect different genes performing equivalent functions in different species, this is inconsistent with a model where the same selective pressure is acting uniformly across species. The disparity in the identity of genes which are imprinted in different species [250] may reflect either de novo imprinting driving speciation [246], de novo imprinting resulting from unique species-specific selective pressure or purifying selection against imprints that are no longer relevant in certain derived species. Molecularly, these adaptations would all manifest as altered regulatory elements, and fixation of ncRNAs. To determine the cause and consequences of imprinting, the most informative data will be systematic functional testing of the consequence of imprinting at each locus, e.g. what effect does activation of the second allele have? A complete catalogue of imprinted genes is essential for this process, and
currently does not exist for any species. Thus in this thesis I asked **to what extent the murine genome is imprinted?** The complete catalogue of imprinted genes may provide additional insight into the function of imprinted genes, particularly if interrogation bias exists in the known subset. Imprinted expression is highly stage and tissue-specific. Distinguishing this level of detail in the patterns of imprinted expression may also offer insight into selective pressure(s) moulding imprinted expression. If a subset of functionally related genes is consistently imprinted in related tissues and biallelically expressed elsewhere, it would suggest that a single selective pressure acts on their role in these tissues. The same logic applies to stage-specific imprinting.
Chapter 2: Oct4 facilitates the expansion of murine posterior progenitors after E7.0

I helped conceive of the project, design the experiments, author the manuscript, and collected the data excluding Fig 7A-C (help from the Toronto Phenogenomics Core) and 9A (contributed by Anthony Lin). The data in 9B,C as well as 10A,B were normalized by Tomas Babak.

This chapter has been submitted for publication:

Abstract

Oct4 is a widely recognized pluripotency factor as it maintains ES cells in a pluripotent state, and, in vivo, prevents the ICM in murine embryos from differentiating into trophectoderm. However its function in somatic tissue after this developmental stage is not well characterized. We set out to investigate the effect of depleting Oct4 in mouse embryos after E7.0, and found that lack of Oct4 at these later stages of development leads to craniorachischisis, random heart tube orientation, failed turning as well as posterior truncation. Oct4 is not required in extraembryonic tissue for these processes, but sufficient expansion of posterior progenitor populations derived from the primitive streak is necessary for completion of these developmental processes and requires embryonic Oct4. Unlike ES cells, separate depletions of the pluripotency factors Sox2 and Oct4 at E7.0 do not phenocopy, suggesting that the function of Oct4 after E7.0 is not within an intact pluripotency network. Instead, the requirement for Oct4 to maintain proliferation, specifically in the primitive streak, offers a single explanation for the multiple embryonic phenotypes seen after Oct4 depletion after E7.0.
Introduction

Oct4 is a homeodomain containing TF of the POU family required for pluripotency in ES cells and preimplantation embryos [252]. It has been extensively characterized in ES cells, and established as a hub of the signalling network that maintains pluripotency along with Sox2, Nanog and others [119-121, 126]. Embryonically, Oct4 is present in the developing mouse zygote and down-regulated somatically between E7.0 and E9.0 depending on the cell type (Fig 1) [139, 142]. After E9.0 of murine development, Oct4 is only present in PGCs and the female gametic lineage where it is required for viability [139, 253]. Oct4 is maintained through the early stages of development by intercellular Nodal acting in part through Smad2 [254, 255].

Conversely, Caudal type homeobox 2 (Cdx2) mediates repression of Oct4 in trophectoderm of the early blastocyst, while both Eomesodermmin (Eomes) and Germ Cell Nuclear Factor mediate repression after implantation in the embryo [256, 257]. Oct4 buffers the ICM against differentiation into trophectoderm, the embryonic contribution to the placenta, but the proposal that \textit{Pou5f1} (gene symbol for Oct4) emergence relates to evolution of the mammalian placenta [258] is not supported given that \textit{Pou5f1} evolved before the origin of amniotes [259]. It is unknown whether Oct4 has any post-implantation function in murine somatic development.

Various roles have been ascribed to homologs, but it is unknown whether any of these post-implantation functions are conserved. \textit{Pou2}, the homolog of Oct4 in \textit{D. rerio}, has been implicated in a variety of processes including: establishing the midbrain-hindbrain boundary [260], embryonic expansion prior to gastrulation [261], hindbrain segmentation [262], and endoderm specification [263]. Maternal deposit \textit{Pou2} also regulates dorsal-ventral axis formation [264]. In \textit{X. laevis}, the Oct4 homolog \textit{XIPou91} has been implicated in mesoderm specification and/or repression of endoderm, neural and organizer specification [265]. Furthermore, \textit{XIPou91}
Figure 1. Oct4 localization from E6.5-9.25 of murine development, based on [142].

is required for posterior elongation and complete anterior development (neural segmentation and eye development) [265]. Aside from maintaining the viability of PGCs, the role of Oct4 in post-implantation murine development is uncharacterized [119, 252, 253, 266].

Depletion of Oct4 from ES cells recapitulates loss from the ICM in that both differentiate into trophectoderm [252], but it is not clear whether inferences gleaned in vitro about Oct4’s role
in post-implantation development translate in vivo. Depending on the culture conditions, over-expression of Oct4 in ES cells promotes primitive endoderm and mesoderm [267], cardiomyocytes [144] or neuronal differentiation [143]. Differentiation signals reciprocally regulate Oct4 and Sox2 protein abundance in an ES cell model of germ layer fate selection, where Oct4 activity promotes mesendoderm at the expense of neural differentiation and Sox2 does the opposite [268]. Germ layer induction begins at E6.0 in vivo, and when it is complete the majority of stem and progenitor cells produced generate progeny restricted to one germ layer. There are exceptions though -- neural crest cells that form the ‘fourth germ layer’ produce daughters of all of the other germ layers, and bipotent ectoderm/mesoderm progenitors persist in the tailbud after E7.5 [269]. However, Oct4 levels may not impact these cell fate decisions in vivo at physiological levels: shuttling kinetics of Oct4 between the nucleus and cytoplasm have proven a better predictor of preimplantation cell fate decisions than Oct4 levels in embryos [270]. Since shuttling kinetics are inferred to be a readout of accessibility of TFs to genomic binding sites and predict preimplantation cell fate outcomes, this suggests that at physiological levels features of chromatin dictating whether Oct4 can bind its genomic targets, and not the levels of Oct4, mediate these early cell fate outcomes. in vivo experiments in adult tissues suggest a comparable distinction: ectopic Oct4 in adult mice induces hyperproliferation but not ectopic differentiation [271]. Contrasting conclusions from in vitro and in vivo approaches leave the function of Oct4 unclear during post-implantation embryonic lineage specification. 

The molecular features distinguishing Oct4’s context-specific impact on cell function also are unclear. Physical interactions suggest Oct4 may have roles in chromatin modification, regulation of transcription, DNA replication and DNA repair as well as post-transcriptional modification, ubiquitination and various other functions [119-121, 272, 273]. Oct4 both activates
and represses transcription [139]. It binds thousands of sites in the ES cell genome, often co-occupying these sites with Sox2, Nanog, Smad1 and Signal Transducer and Activator of Transcription 3 (Stat3) [126]. The majority of genes occupied by several of these TFs are active in ES cells, but their binding does not ensure expression [126].

Somatic tissues after E7.0 are not considered pluripotent, so the presence of Oct4 until E9.0 suggests it may have a different function in this window of development than in the ICM. Given that Pou5f1 null embryos arrest development at E3.5, we asked what role Oct4 had in murine development after E7.0 using a conditional system. In vitro studies have inferred many roles for Oct4 at this stage of development, including regulating neural versus mesendoderm differentiation [268] as well as promoting cardiomyocyte [144] and neuronal differentiation [143], but it is not clear whether these capture its role in vivo during post-implantation development. We show that Oct4 removal after E7.0 results in craniorachischisis, random heart tube orientation, failed turning as well as posterior truncation. We did not observe reciprocal diversion between germ layers, from mesoderm to neuroepithelium after depletion of Oct4 and from neuroepithelium to mesoderm after depletion of Sox2, as inferred from in vitro models. As such we report that diversion of progenitor populations is not responsible for the phenotype. The phenotype is not the result of a general delay in development, nor is it a general failure of the pluripotency network. Depletion of Sox2, another core member of the pluripotency network, in an overlapping window of development does not phenocopy Oct4 depletion. Instead, Oct4 is required after E7.0 to maintain proliferation in the primitive streak.
Materials and Methods

Animal Husbandry

All procedures were approved by the University of Toronto Animal Care Committee in accordance with the Canadian Council on Animal Care. For staging, embryos were assumed to be 0.5 dpc at noon on the day a vaginal plug was found. The following stocks were used in the study: CD1 (Charles River), Oct4\textsuperscript{f/f} [253], Z/EG [274], B6.Cg-Tg(Hist1H2BB/Egfp)1Pa/J [275], Bry-Cre [276], Sox1-Cre [277], Foxa2\textsuperscript{tm2.1(cre/Esr1*)Moon}/J [278], Sox2\textsuperscript{f} [129], CreER\textsuperscript{T2} [279]. Individual embryos or the associated extraembryonic tissues were genotyped as originally described.

Tamoxifen Administration

99mg of tamoxifen (Sigma) was dissolved by sonication in a solution of 100ul of ethanol (Sigma) and 1ml of peanut seed oil (Sigma). The solution was kept in a ~55°C waterbath during preparation and prior to administration to avoid precipitation. 50µl doses of this solution were administered to pregnant mothers by oral gavage using a 250µl gastight #1725 syringe (Hamilton). Tamoxifen was administered at the timepoint(s) indicated at the end of each figure caption.

Statistical Analysis

The density of mesenchyme, frequency of apoptosis and proliferation, relative abundance of transcripts and thickness of neuroepithelium were compared using 2-way ANOVAs. F-values from the embryonic genotype’s contribution (Oct4\textsuperscript{f/f} vs Oct4\textsuperscript{f/f};CreER\textsuperscript{T2+/-}) to variation are indicated except for Fig 4H and 11K where the embryonic segment’s contribution is reported.
Binding enrichment amongst differentially expressed genes was compared using hypergeometric tests. The thickness of notochords was compared using a two-tailed t-test. A threshold of p<0.05 was used for each test (ANOVA, hypergeometric and t-test).

Quantitative PCR

QPCR was performed using Taqman Assays (Applied Biosystems) on a 7900 HT Fast-Time PCR System (Applied Biosystems) and analyzed with software provided by the manufacturer: SDS 2.3. The following Taqman Assays were used: *Pou5f1* Mm00658129_gH, *Hprt* Mm00446968_m1, *Gusb* Mm01197698_m1, *Gusb* Mm01197698_m1, *Rplp0* Mm99999223_gH, *Sox2* Mm03053810_s1, *Smad1* Mm00484723_m1, *Cdx2* Mm00432449_m1, *Klf2* Mm01244979_g1, *Eomes* Mm01351984_m1, *Lefty2* Mm00774547_m1, *Dll1* Mm01279265_g1, *Bax* Mm01205547_g1, *Bcl2* Mm00477631 and *Cdkn1c* Mm01272135_g1.

Generating Chimeras

Chimeras were produced as outlined in [280]. In brief, females were superovulated with 5-10 IU Folligon (Intervet) and 5 IU Chorulon (Intervet), embryos flushed using M2 (Millipore), zona pellucididas removed using dose Tyrode’s Acidic Solution (Sigma) and morula aggregated overnight in KSOM +AA with D-Glucose (Millipore) under mineral oil (Millipore). Aggregates were reimplanted into the uterus of pseudopregnant females the following morning, which was considered E2.5. Contribution was assessed by semi-quantitative PCR. For tetraploid chimeras, 4-cell embryos were electrofused at 30V for 40usec, aggregated with RFP ES cells [281] once they reached the 8-cell stage, and then transferred into pseudopregnant females the following morning. To control for timing of development and induction in both the diploid and tetraploid
experiments, embryos of different genotypes were mixed before transferring to surrogates and genotyped following dissection. All surrogate mothers were induced with tamoxifen at E6.0 and E6.5.

**Immunohistochemistry and Histology**

Oct4 staining was performed as described previously [142]. For all other immunohistochemistry, embryos were fixed in 4% PFA overnight at 4°C, sectioned at a thickness of 10µm and primary antibodies applied overnight at 4°C at the following concentrations: Chordin 1:100 (R & D Systems), p-Smad1 1:400 (Cell Signalling), Caspase-3 1:500 (Promega), Ph3 1:500 (Cell Signalling), Cdx2 1:100 (Biogenex), Bry 1:50 (R & D Systems), Sox2 1:50 (R & D Systems). An antigen retrieval step of boiling the sample in 10mM Sodium Citrate Buffer, pH 6.0 for 15 min was used for Chordin staining. Hematoxylin and Eosin (Sigma) staining was performed according to the manufacturer’s instructions.

**Microarray and Statistical Enrichment Analysis**

Embryos and extraembryonic tissue were dissected, marked as pairs and immediately flash frozen separately in liquid nitrogen. The extraembryonic tissues were used to infer the embryonic genotype. Embryos were then pooled according to their genotype, i.e. those with CreER<sup>T2</sup> separated from those without for each timepoint. All samples at each timepoint had at least two embryos of either genotype. RNA was extracted with Trizol according to the manufacturer’s instructions (Invitrogen) and sent to the UHN Microarray Centre (Toronto, ON, Canada) for fluor-labeling (protocol GE2 v5.7), microarray hybridization, and array scanning. In brief, Cy3 and Cy5 (NEN Life Science) were incorporated into cDNA using Superscript II (Invitrogen) and
fluor-labelled cDNA was purified using CyScribe GFX purification kit (GE). Hybridizations were conducted in DIG Easy Hyb solution (Roche) using fluor-reversed pairs (i.e. each sample was hybridized twice to two different arrays labelled once with Cy3 and once with Cy5) to Agilent 44k Mouse Gene 1.0ST Gene Chips, scanned using G2565C DNA Scanner, and features extracted using Agilent Feature Extraction Software as described in the UHN Microarray Centre’s version 10.5 (protocol GE2_105_Dec08). Intensities were averaged across fluor-flips prior to normalization. Data were normalized [282] and clustered using Rosetta Resolver. Enrichment for TF binding within manually identified clusters was assessed for previously defined genome binding coordinates [126] where a gene with transcription start site within 5 kB of a binding site was considered bound. Differential expression was defined using a p-value threshold of 0.01 (i.e. 1% of genes exceed this threshold in technical replicate hybridizations) and background comprised of genes denoted "present" by the software. For pathway enrichment, the following databases were used to define functional categories: GeneGo Pathways, Reactome, Kegg, NCI, Ingenuity Pathways, Biobase, and Biocarta.

**Results**

**Embryonic development requires Oct4 after E7.0**

To assess the role of Oct4 after E3.5 we used floxed alleles of Oct4 [253] and a tamoxifen inducible recombinase (CreER<sup>T2</sup>) expressed ubiquitously from the ROSA locus [279]. Oct4 protein is depleted ~20 hours after tamoxifen induction (ATI) at E6.875 (approximately E7.75 in this scheme (Fig 2A-D)). Embryos induced at E7.875 and analyzed at E9.5 were phenotypically wild-type (WT) indicating that Oct4 is not required for development after this
stage (3 litters). Conversely, embryos induced at E5.875 and analyzed at E9.5 all had severe phenotypes (3 litters), where the embryonic anterior was distinguishable, but amorphous, and the embryonic posterior was absent (Fig. 3A,B). E9.5 embryos in which excision of Pou5f1 was initiated at E6.875 failed to turn, had severe posterior truncations, randomly oriented heart tubes, craniorachischisis (open neural tube along its entire length) as well as impaired somitogenesis (Fig. 3C-E). Such animals will be referred to as Oct4COND MUT throughout. The Oct4COND MUT phenotype is fully penetrant in embryos treated at E6.875 and analyzed at E9.5 (n=63). The phenotype is not a consequence of tamoxifen itself, leaky recombinase activity prior to induction, or associated with recombination of a single Pou5f1 allele: Oct4flox/flox(fl) embryos induced at E6.875, uninduced Oct4flox/fl;CreERT2+/− embryos, and Oct4flox/fl;CreERT2+/− embryos induced at E6.875 are all (n>3 litters each) phenotypically WT at E9.5. Rather, the phenotype is due to Oct4 depletion after E7.0.
Figure 3. The Oct4\textsuperscript{COND MUT} phenotype. A,B Oct4\textsuperscript{f/f} (A) and Oct4\textsuperscript{f/f};CreER\textsuperscript{T2+/−} (B) embryos after tamoxifen induction (ATI) at E5.875 and dissection at E9.5. The phenotype of the Oct4\textsuperscript{f/f};CreER\textsuperscript{T2+/−} embryo is severe compared to induction at either E6.875 or E7.875, as the embryonic posterior is absent and anterior structures are poorly formed. The images are of a freshly dissected embryo, with the posterior and anterior separated by a dashed line. A control is provided for comparison in panel ‘E.’ Heart tube (*), posterior (#), pharyngeal arch (^) and forebrain (%) are indicated.

C-E Oct4\textsuperscript{f/f};CreER\textsuperscript{T2+/−} phenotype ATI at E6.875 and dissection at E9.5. Oct4\textsuperscript{f/f} are phenotypically WT littermates of Oct4\textsuperscript{COND MUT} (Oct4\textsuperscript{f/f};CreER\textsuperscript{T2+/−}). C Oct4\textsuperscript{f/f} (left), Oct4\textsuperscript{COND MUT} (right). Arrows indicate somites that are absent in the Oct4\textsuperscript{COND MUT} embryos. D Dorsal view of Oct4\textsuperscript{f/f};CreER\textsuperscript{T2+/−} littermates. Arrows indicate open neural tube that is present in all Oct4\textsuperscript{COND MUT} embryos. E Oct4\textsuperscript{f/f} (WT) heart tube orientation (left) and Situs Inversus orientation (right) are outlined with dashed red lines in two Oct4\textsuperscript{f/f};CreER\textsuperscript{T2+/−} littermates. F The percentage of CreER\textsuperscript{T2+} embryos recovered at each developmental stage. The decrease indicates that Oct4\textsuperscript{COND MUT} resorb after E9.5. G,H Representative images of an E10.5 Oct4\textsuperscript{f/f};CreER\textsuperscript{T2+/−} embryo and Oct4\textsuperscript{f/f} littermate. Features such as the otic cup (^) and forelimb bud (#) which arise in WT development after neural tube closure is initiated and turning is complete are present in both Oct4\textsuperscript{f/f};CreER\textsuperscript{T2+/−} and Oct4\textsuperscript{f/f} littermates. Conversely, defects in turning, somitogenesis, neural tube closure and posterior extension all fail to recover in Oct4\textsuperscript{f/f};CreER\textsuperscript{T2+/−} embryos by E10.5. G Sagittal view. H Dorsal view.

Oct4 depletion does not cause a general delay in development. Oct4\textsuperscript{f/f};CreER\textsuperscript{T2+/−} embryos were recovered 1:1 with Oct4\textsuperscript{f/f} littermates until E9.5, but less frequent recovery at later stages of development indicates failure to develop past E9.5 (Fig. 3F). Affected aspects of development lag behind unaffected features, which is inconsistent with a general delay in development (Fig. 3).
Indentation of the otic cup occurred and the branchial arches formed in Oct4\textsuperscript{COND MUT}, events that normally occur by E9.0. Forelimb buds also protruded in Oct4\textsuperscript{COND MUT} as they do in WT embryos by E9.5. Conversely, while the neural tube closes rostrally between E8-9 and caudally by E9-10 [283], turning occurs by ~9.0 and posterior extension reaches 21-29 somites by E9.5 in WT embryos. These events always failed at E9.5 when \textit{Pou5f1} excision was induced at E6.875 (26.5 vs 4.6 somites; Oct4\textsuperscript{f/f} vs Oct4\textsuperscript{f/f};\textit{CreER}\textsuperscript{T2+} littermates; n=63; Fig. 2C).

Additionally, heart tube orientation was randomized, 38.6\% of Oct4\textsuperscript{f/f};\textit{CreER}\textsuperscript{T2+} had situs inversus and 6.8\% had situs ambiguous (p>0.05 Chi-square test; Fig. 4A). Mesenchyme density was also reduced in Oct4\textsuperscript{COND MUT} (F\textsubscript{1,13}=54.59, p<0.05 2-way ANOVA; *p<0.05, **p<0.001 Bonferroni posttest; Fig. 4B-E). The neuroepithelium of Oct4\textsuperscript{COND MUT} embryos was also thicker in regions, particularly in the distal portion of the embryo (F\textsubscript{1,287}=94.95, p<0.05 2-way ANOVA; ***p<0.001 Bonferroni posttest; Fig. 4F-H). Consistent with the absence of a general delay in development, the features disrupted in Oct4\textsuperscript{COND MUT} were still arrested while others had developed in the rare mutants which persisted beyond E9.5 (Fig. 3G,H). These data indicate that Oct4 is required for posterior extension, turning, heart tube orientation and neural tube closure (NTC).
Figure 4. Characterization of the Oct4\textsuperscript{COND MUT} phenotype. A Quantification of the frequency heart tube orientation is altered in Oct4\textsuperscript{f/f,CreER\textsuperscript{T2+/+}} embryos (n = 44). B,C Representative Hematoxylin and Eosin (H and E) sections from the trunk (see ‘Tr’ in panel E) at E9.5. Differences in the density of mesenchyme, which is indicated with two-headed arrows, are apparent. B Oct4\textsuperscript{f/f} C Oct4\textsuperscript{COND MUT}. D Mesenchyme, connective tissue comprised of mesoderm and neural crest cells, has a lower density in Oct4\textsuperscript{COND MUT}. The average density within 4000\textsuperscript{um\textsuperscript{2}} ± s.e.m. is plotted (F\textsubscript{1,13} = 54.60, p<0.05 2-way ANOVA; *p<0.05, ***p<0.001 Bonferroni posttest). Insufficient mesenchyme was present in the tail of Oct4\textsuperscript{COND MUT} to quantify density due to the posterior truncation. Schematic illustrating where mesenchyme density samples were taken from is provided in panel ‘E.’ E Schematic illustrating the sites of NT closure points, the directions in which the NT ‘zippers’ shut as well as the sections where mesenchyme density was quantified. F,G,H The neuroepithelium of Oct4\textsuperscript{COND MUT} embryos is thicker in part of the distal segment. Generally the region of thick neuroepithelium occurred in close proximity to where the first closure point would normally occur, dorsal to the first few somites. The red box in ‘G’ indicates a thick region of neuroepithelium in a Oct4\textsuperscript{COND MUT} embryo and an equivalent region in a control Oct4\textsuperscript{f/f} embryo is marked in ‘F.’ Both the embryos in ‘F’ and ‘G’ were induced with tamoxifen at E6.875 and E7.375. The two-headed arrow indicates the thickness of
neuroepithelium. H Quantification of the average neuroepithelial thickness, comparing the distal region of Oct4\textsuperscript{COND MUT} embryos to adjacent regions. Error bars are ±s.e.m. The ventral side of each embryo faces left in each case. An ANOVA was used to compare thickness of the distal to adjacent regions within each embryo (F\textsubscript{1,287}=94.95, p<0.05 2-way ANOVA; ***p<0.001 Bonferroni posttest). All scale bars are 50µm.

Failed caudal NTC and posterior extension are related defects, while failed cranial NTC and turning are independent defects

Tamoxifen induction at E7.375 led to incomplete penetrance of the Oct4\textsuperscript{COND MUT} phenotype (Fig. 5A,B), suggesting that, after accounting for system kinetics, ~E8.0 marks the end of the developmental window in which embryos are sensitive to Oct4 depletion. Complete coincidence between pairs of features in litters with incomplete penetrance of the Oct4\textsuperscript{COND MUT} phenotype suggests dependence of these pairs of features. Posterior extension precedes caudal NTC in development, and the complete coincidence of their failures suggests a dependence of NTC caudal to closure point 1 (see Fig. 4E) on posterior extension (n=23; 23/36 embryos induced at E7.375 had caudal NTC defects and these same 23 embryos had posterior truncation defects. Amongst these 36 embryos, none had only one of these two defects). Conversely, embryos that had turning defects and WT posteriors (n=2/13 embryos with WT posteriors) supports independence of these two processes. Likewise, NTC rostral to point 1 failed in some mutants that turned and had WT posteriors (n=9/13 embryos with WT posteriors), indicating independence of the rostral NTC defect from these processes. These data suggest independent requirements for Oct4 after E6.875 in NTC rostral to closure point 1, turning, and posterior extension/caudal NTC.
Figure 5. Characterization of the Oct4<sup>COND MUT</sup> phenotype continued. A,B Penetrance declines with later doses of tamoxifen. A Representative Oct4<sup>fl/fl</sup>;CreER<sup>T2+/−</sup> litter at E9.5 exhibiting incomplete penetrance of mutant features when tamoxifen is administered at E7.375 and E7.875. Heart tube (*), posterior (#), pharyngeal arch (^) and forebrain (%). Within the litter, ‘i’ is phenotypically WT, ‘ii’ has an open NT between closure points 1 and 2 that is marked with dashed line, and ‘iii’ has the defects characteristic of Oct4<sup>COND MUT</sup> embryos: a truncated posterior where somites have not formed (compare ‘ii’ # to ‘iii’ #), that has not turned (note how ‘ii’ faces its tail, but ‘iii’ does not), and an open NT along its entire length marked with an unbroken line (note the ridge of unclosed neuroepithelium in the bracketed region of ‘ii’ or ‘iii’ as compared to ‘i’ where the NT is closed and the ridge is absent. B Quantification of mutant features in embryos induced at E7.375 and E7.875. Note that ‘defective’ neural tube closure indicates craniorachischisis, while ‘wild type’ denotes a neural tube that is closed posterior to closure point 1 and open anterior to it (also a neural tube closure defect, albeit less severe than craniorachischisis). C,D,E Cross-sections of Oct4<sup>COND MUT</sup> notochords contain more cells. All litters (C-E) were induced with tamoxifen at E6.875 and E7.375. C A transverse section of a phenotypically WT Oct4<sup>fl/fl</sup> E9.5 embryo, with the notochord outlined with a red box. D A Oct4<sup>COND MUT</sup> embryo with the notochord outlined with a red box E Quantification of the average notochord thickness throughout embryos (cells/cross-section): Oct4<sup>fl/fl</sup> compared to Oct4<sup>COND MUT</sup> (two-tailed t-test, *p<0.05). All scale bars are 50µm.
Convergence and extension movements elongate the embryo in the anterior-posterior axis bringing the neural folds into opposition prior to adhesion at closure point 1 (see Fig. 4E). Characterized mutants that fail to close at closure point 1 (craniorachischisis) are from the planar cell polarity (PCP) pathway [283, 284], which mediates convergence and extension as well as left-right asymmetry [285]. Notochord diameter, which narrows during extension, is enlarged in Oct4COND MUT (p<0.05, two-tailed student t-test; Fig. 5C-E), consistent with deficient extension. Thus disruption of PCP signalling may contribute to craniorachischisis and left-right asymmetry defects (randomized heart tube orientation) characteristic of these mutants.

NTC rostral and caudal to closure point 1 occur by different mechanisms. Contraction of actin-myosin microfilaments is crucial in closure of cranial, but not spinal neuroepithelium [283]. In the intermediate spine, bending occurs at the median hinge point (MHP) and dorsolateral hinge points (DLHPs) to oppose the neural folds. At the lower end of the spine only DLHPs are used, while in the upper spine DLHPs are inhibited by factors released from the notochord and MHPs oppose the neural folds [283]. Unlike the spinal region where expansion of paraxial mesoderm is not required for elevation and subsequent NTC, cranial NTC is initiated by expansion of underlying mesenchyme [283]. The density of cranial mesenchyme was lower in Oct4COND MUT (Fig. 4D), indicating that Oct4 is required for expansion of cranial mesenchyme that is required for cranial NTC.

Extraembryonic Oct4 is not required after E7.0, and embryonic loss can be compensated by WT cells
To distinguish autonomous from non-autonomous effects, we tested whether Oct4 removal in select tissues affected development of others. Oct4 is present in the neuroepithelium until ~E8.5, portions of definitive endoderm until E9.0 and portions of mesoderm until ~E8.25 with the final positive components of mesoderm being the node and paraxial mesoderm [142]. Oct4 was removed in the neuroepithelium using Sox1-Cre [277], which is expressed and catalytically active from E7.5; in definitive endoderm using tamoxifen-inducible Foxa2<sub>mcm</sub> [278], which is expressed in the primitive streak at E6.25 and later in the notochord and the floor plate; as well as in embryonic mesoderm using Brachyury (Bry)-Cre [276], which is expressed and catalytically active from E6.25. In each case, Oct<sup>f/f</sup>;Z/EG<sup>+/+</sup> females were mated with Oct<sup>f/+</sup>;lineage-specific (Sox1, Bry, Foxa2) Cre<sup>+/+</sup> males and the progeny analyzed at E9.5. Recombination at the Z/EG locus yields GFP expression [274]. In this scheme, Oct<sup>f/+</sup>;Z/EG<sup>+/+</sup>;lineage-specific Cre<sup>+/+</sup> embryos should reveal aspects of the Oct4<sup>CND MUT</sup> phenotype related to specific requirements for Oct4 within each lineage tested or potentially abort by E9.5 if the phenotype is more severe than with CreER<sup>T2</sup>. In all cases embryos emerged in the anticipated ratios with no phenotype: Oct<sup>f/+</sup>;Z/EG<sup>+/+</sup>;lineage-specific Cre<sup>+/+</sup> embryos were present and not under-represented (1:4 anticipated based on breeding scheme) without any aspects of the Oct4<sup>CND MUT</sup> phenotype (n=17/62, using Sox1-Cre, n=17/39 using Bry-Cre, and n=14/55 using Foxa2<sub>mcm</sub>). These data suggest that Oct4 is not required within domains where Sox1, Bry or Foxa2 are present for development. Since false-negatives may arise due to low recombination efficiency in this scheme, we used the GFP expression resulting from recombination at the Z/EG locus in Oct<sup>f/+</sup>; Z/EG<sup>+/+</sup>;lineage-specific Cre<sup>+/+</sup> embryos as a proxy for efficiency of each lineage-specific Cre. While Sox1-Cre and Bry1-Cre induced considerable recombination within their respective domains (>95% and >51%, respectively; Fig. 6A-C), Foxa2<sub>mcm</sub> yielded
negligible recombination by E9.0 (<5%; data not shown), so we did not use it in additional experiments. To test whether Sox1-Cre and Bry-Cre did not recapitulate the Oct4\textsuperscript{COND MUT} phenotype because mono-allelic rather than biallelic recombination was prevalent at the Oct4\textsuperscript{f/f} locus, wherein viability of Oct4\textsuperscript{+/−} embryos indicates that sufficient Oct4 would be transcribed from the intact allele, we assessed the development of embryos where one Pou5f1 allele was removed prior to introduction of the recombinases. Even with this sensitized approach, progeny at E9.5 were produced in anticipated ratios (1:4) and all phenotypically WT, suggesting Oct4 is not required after expression from either the Sox1 or Bry locus is initiated (n=7/20 Oct4\textsuperscript{∆};Z/EG\textsuperscript{+/-};Sox1-Cre\textsuperscript{+/-}; n=7/27 Oct4\textsuperscript{∆};Z/EG\textsuperscript{+/-};Bry-Cre).

To assess whether Oct4 depletion influenced cell fate, which does not require efficient recombination, we compared the distribution of GFP\textsuperscript{+} cells in embryos where recombinase activity removed Pou5f1 and activated GFP (Oct4\textsuperscript{∆};Z/EG\textsuperscript{+/-};lineage-specific Cre\textsuperscript{+/-}) to control embryos where recombinase activity only activated GFP (Oct4\textsuperscript{∆};Z/EG\textsuperscript{+/-};lineage-specific Cre\textsuperscript{+/-}). Based on the distribution of GFP\textsuperscript{+} cells being unaffected by Oct4 depletion, we infer that cell fate does not change following depletion in these expression domains. This experimental scheme cannot assess whether Oct4 influences cell fate during the developmental window when lineages are specified to express Sox1 or Bry, but have
Figure 6. Timing of Oct4 depletion is relevant to penetrance of the $Oct4^{COND\text{ MUT}}$ phenotype, but diversion of cells between germ layers is not. A-C Lineage specific Cre induce lower rates of recombination than CreER$^{T2}$ during the developmental window in which embryos are sensitive to Oct4 loss. A Sagittal section of E9.0 Sox1-Cre$^{+/+};Z/EG^{+/+}$ embryo. B Sagittal section of E9.0 Bry-Cre$^{+/+};Z/EG^{+/+}$ embryo. The sections (A,B) are oriented with the ventral side facing left. C Quantification of recombination frequency at E7.75 and E9.0, comparing Sox1-Cre and Bry-Cre. D,E Oct4 depletion in the Bry domain does not result in cellular diversion to neuroepithelium. D Sagittal sections of E9.5 embryos wherein one allele of Oct4 has been removed and the other is intact ($Oct4^{\Delta/+};Z/EG^{+/+};Bry-Cre^{+/+}$). GFP marks cells where recombination has occurred. E Sagittal section of E9.5 embryos wherein Oct4 has been depleted ($Oct4^{\Delta/+};Z/EG^{+/+};Bry-Cre^{+/+}$). The number of Bry$^+$ cells are present in the neuroepithelium of E9.5 embryos is comparable to controls ‘D’ which does not support diversion of cells into the neuroepithelium following Oct4 depletion. GFP marks cells where recombination has occurred. Magnified insets in the upper right correspond to the section outlined in each panel (D,E). Scale bars are 200µm (D,E).

yet to deplete Oct4 due to system kinetics. Aside from this caveat, these data suggest that Oct4 loss does not result in diversion of mesoderm to neural tissue or vice-versa (Fig. 6D,E).
An alternate explanation for the Oct4\textsuperscript{COND MUT} phenotype could be a requirement for Oct4 in extraembryonic tissue: Pou5f1 is expressed in extraembryonic mesoderm until E8.0, allantoic angioblasts until E9.0 [142] as well as extraembryonic endoderm which promotes proliferation and organization of the primitive streak until E8.0 [286]. To test this possibility, Red fluorescent protein positive (RFP\textsuperscript{+}) ES cells were aggregated with tetraploid Oct4\textsuperscript{f/f};Z/EG\textsuperscript{+/-};CreER\textsuperscript{T2+/} embryos, where ES cells contribute to the embryo, and tetraploid cells generate all extraembryonic lineages other than fetal blood, fetal vessels and the yolk sac endoderm [287]. In this scheme, induction with tamoxifen will result in selective removal of Oct4 from the tetraploid extraembryonic lineages. In contrast to Oct4\textsuperscript{f/f};Z/EG\textsuperscript{+/-};CreER\textsuperscript{T2+/} positive controls which exhibited the characteristic Oct4\textsuperscript{COND MUT} phenotype, tetraploid Oct4\textsuperscript{f/f};Z/EG\textsuperscript{+/-};CreER\textsuperscript{T2+/} embryos induced at E6.375 and E6.875 supported development of WT ES-derived embryos to E9.5 (Fig. 7A-C). Embryos were dosed on this relatively early schedule to avoid false negatives that might result from the delay in development associated with transferring embryos to pseudopregnant mothers. These data demonstrate that the Oct4\textsuperscript{COND MUT} phenotype is due to a requirement for Oct4 in embryonic tissue.

Recombination frequency and cell non-autonomous effects might influence phenotype penetrance in the experiments using lineage-specific recombinases. To investigate their influence, we generated diploid chimeras by aggregating WT and Oct4\textsuperscript{f/f};HisGFP\textsuperscript{+/-};CreER\textsuperscript{T2+/} morulas to determine whether the deficiency is cell
Figure 7. Oct4 is required in the embryo after E7.0, but not in extraembryonic tissue (see also Figure 3). All litters (A-E) were induced with tamoxifen at E6.0 and E6.5 to compensate for the variability in developmental timing associated with transferring embryos. A A representative embryo when Oct4+/+ RFP ES cells are aggregated with a tetraploid Oct4f/+ embryo. The control aggregation yields an E9.5 chimeric embryo that is phenotypically WT with Oct4f/+ extraembryonic tissue. B A representative positive control. An Oct4f/+;CreERT2+/− E9.5 embryo with the Oct4COND MUT phenotype. C Oct4 depletion in extraembryonic tissue after E7.0 is compatible with WT development. A representative chimera consisting of RFP+ Oct4+/+ ES cell derived embryo and tetraploid Oct4f/+;CreERT2+/− extraembryonic tissue. The embryo has turned (compare panel ‘B’ where the tail is behind to panel ‘C’ where it is in front), undergone NTC and posterior extension (compare the lack of somites and short tail in panel ‘B’ to the somites and full-length tail in ‘C’). D A representative image of the most severe defect observed in a diploid chimeras consisting of Oct4f/+ and Oct4f/+;CreERT2+/− cells. The neural tube is open between closure points 1 and 2, indicated here with a black bracket. All Oct4COND MUT features aside from cranial NTC defect, which is still present in 5/16 mosaic embryos, are rescued by Oct4+/+ cells in these diploid chimeras (16/16). For example, this embryo has ‘turned’ such that it faces its tail and the posterior has extended to WT length. All scale bars (A-D) are 200 um. E Quantification of genotypes and phenotypes of recovered chimeric embryos.

<table>
<thead>
<tr>
<th>Group</th>
<th>Tetraploid Chimeras</th>
<th>Diploid Chimera</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Negative Control (panel A)</td>
<td>Positive Control (panel B)</td>
</tr>
<tr>
<td>Genotype of diploid morula 1</td>
<td>Oct4f/+; CreERT2+/−</td>
<td>Oct4f/+; CreERT2+/−</td>
</tr>
<tr>
<td>Genotype of diploid morula 2</td>
<td>Oct4f/+</td>
<td>Oct4f/+ (CD1)</td>
</tr>
<tr>
<td>Genotype of tetraploid morula</td>
<td>Oct4f/+</td>
<td>Oct4f/+; CreERT2+/−</td>
</tr>
<tr>
<td>Genotype of ES Cells</td>
<td>RFP+;Oct4+/+ (129S6/B6)</td>
<td>RFP+;Oct4+/+ (129S6/B6)</td>
</tr>
<tr>
<td>n</td>
<td>6/6</td>
<td>5/5</td>
</tr>
<tr>
<td>Phenotype</td>
<td>WT</td>
<td>Oct4COND MUT</td>
</tr>
</tbody>
</table>
autonomous. After tamoxifen induction at E6.375 and E6.875, we observed complete rescue in 11/16 embryos where Oct4 was depleted in a mosaic fashion, and only rostral NTC deficit in the other 5/16 chimeras (Fig. 7D,E). Contribution by Oct4\(^{ff}\);HisGFP\(^{+/-}\);CreER\(^{T2+/-}\) morulas ranged from 20-60% in these 16 chimeras. This result indicates that WT cells compensate for Oct4\(^{+/-}\) cells either via cellular proliferation or by release of non-autonomous factors. This result also indicates that mosaic Oct4 depletion is insufficient to produce the Oct4\(^{COND MUT}\) phenotype.

Next, we assessed whether the frequency of Sox1-Cre and Bry-Cre induced recombination is sufficient during the developmental window when embryos are sensitive to Oct4 depletion to assay whether a primary defect in one tissue produces other non-autonomous aspects of the Oct4\(^{COND MUT}\) phenotype as secondary effects. By E9.0 Sox1-Cre (>95%) and Bry-Cre (>51%) yielded considerable recombination in their respective domains (Fig. 6A-C), however prior to E8.0, when embryos are sensitive to Oct4 loss, both yielded inefficient recombination (<5%; Fig. 6C). Coupled with the diploid aggregation data that WT cells can compensate for Oct4\(^{+/-}\) cells when these populations are mosaic, the recombination efficiency of both Sox1-Cre or Bry-Cre before E8.0 is insufficient to recapitulate the Oct4\(^{COND MUT}\) phenotype. Because the recombination efficiency of Sox1-Cre and Bry-Cre coincident with the onset of their expression is low, this paradigm does not resolve whether Oct4 is required ubiquitously in lineage specified progenitors between E6.25 and E9.0 or in a specific population such as Oct4\(^+\)Sox1\(^+\) neuroepithelium or Oct4\(^+\)Bry\(^+\) mesoderm. Taken together, these data indicate that the Oct4 deficiency is either cell non-autonomous or can be rescued by cellular compensation from Oct4\(^{+/-}\) cells. These data indicate that Oct4 is required in some capacity after E7.5, but not extra-embryonically or within lineage specified populations.
The functions of Oct4 and Sox2 do not overlap after E7.0

To test whether disruption of the pluripotency network causes the Oct4<sup>COND MUT</sup> phenotype, we removed Sox2 using the same conditional approach [129]. Sox2 is a core component of the pluripotency network that complexes with Oct4, co-occupies many genomic sites and is required for maintenance of Pou5f1 expression in ES cells. ES cells differentiate into trophectoderm when Sox2 is removed [124], however the ability of Oct4 over-expression to rescue pluripotency in these cells suggests that the critical role of Sox2 in pluripotency is to maintain Pou5f1 expression [124]. Sox2 null embryos lack epithelial cells typical of the epiblast and have a later extraembryonic defect which does not permit development past E7.5 [288].

When tamoxifen was administered at E6.375 and E6.875 to Sox2<sup>f/f</sup>;CreER<sup>T2+/-</sup> embryos [129], fluid accumulated in the neural tube of 11/20 Sox2<sup>f/f</sup>;CreER<sup>T2+/-</sup> and 2/20 others had crooked spines at E9.5 (Fig. 8). Thus Sox2 depletion did not phenocopy Oct4 depletion after E7.0. These data do not rule out partial compensation for loss of Sox2 by redundant factors, however Oct4 and Sox2 only partially overlap spatially after E6.0 (prior to Oct4 depletion there is complete overlap in neuroepithelium, considerable overlap in endoderm and negligible overlap in the primitive streak and mesoderm) [142, 288] and the distinct phenotypes produced by excision of Sox2 and Pou5f1 ~E7.0 support the functions of one or both being distinct and non-overlapping after E7.0, in contrast to their roles in ES cells.
Figure 8. Depletion of core pluripotency network factor Sox2 does not phenocopy Oct4 depletion. All litters (A-C) were induced with tamoxifen at E6.375 and E6.875. A 11/20 Sox2^{fl/fl};CreER^{T2+/−} have bloated heads (top row) compared to phenotypically WT Sox2^{fl/fl} littermate controls (bottom row) at E9.5. B Dorsal view of Sox2^{fl/fl};CreER^{T2+/−} (left) with bloated head and Sox2^{fl/fl} littermate (right). Two-headed arrows indicate region where neuroepithelium does not approach the midline of Sox2^{fl/fl};CreER^{T2+/−} as it does in phenotypically WT Sox2^{fl/fl} embryos. C 2/20 Sox2^{fl/fl};CreER^{T2+/−} had crooked spines (indicated with arrows) without bloating (at left and right), while the embryo in the middle has the more prevalent bloated head. No bloated heads or kinked spines were observed in Oct4^{COND MUT} embryos. Conversely, none of the features disrupted in Oct4^{COND MUT} were observed in Sox2^{fl/fl};CreER^{T2+/−}.

Oct4 acts as a repressor, and at Sox2 target sites, an activator, after E7.0

Oct4 binds 3761 genomic loci in ES cells [126]; to determine which targets might be contributing to the Oct4^{COND MUT} phenotype, we measured gene expression changes that occurred after Oct4 depletion. Oct4 protein abundance was similar 15 hours ATI, but degradation became pronounced 20 hours ATI (Fig. 2A-D). To assay early expression changes after Oct4 depletion, RNA was extracted 24, 36 and 48hrs ATI for analysis. Oct4^{fl/fl};CreER^{T2+/−} embryos were separated from Oct4^{fl/fl} littermates for comparison by genotyping extraembryonic tissue, and a reduction of Oct4 transcript was greater than 95% when measured in embryos with CreER^{T2} at each time-point (Fig. 9A).
Many genes already associated with Oct4 in studies with ES cells were differentially expressed when Oct4 depletion was induced at E6.875. For example, *Xist*, which is directly repressed by Oct4 in ES cells was among the most up-regulated genes 24 and 36 hours ATI, supporting Oct4-mediated repression of *Xist* after E7.5 [289]. Other genes already linked to Oct4 which ranked amongst the most up-regulated include *Acin1, Brwd1, Cabin1, Chd1, Chd3, Chd4, Emd, Ino80, Kpna3, Rest, Rif1, Smarca4, Smarca5, Sp1, Top2a* and the most depleted include *Chd5, Cubn, Dppa5a, Fgf8, Hdac2, Hnrnpab, Klf2, Lefty1, Nodal* and *Wnt3a* [119, 122, 123, 290-295]. Enrichment for genomic targets of Oct4 is expected with this approach, but many

<table>
<thead>
<tr>
<th>Upregulated Genes</th>
<th>24 Hrs ATI</th>
<th>36 Hrs ATI</th>
<th>48 Hrs ATI</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF</td>
<td>p-value</td>
<td>TF</td>
<td>p-value</td>
</tr>
<tr>
<td>c-Myc</td>
<td>1.63E-10</td>
<td>c-Myc</td>
<td>2.78E-05</td>
</tr>
<tr>
<td>n-Myc</td>
<td>2.95E-04</td>
<td>Smad1</td>
<td>4.45E-03</td>
</tr>
<tr>
<td>Sox2</td>
<td>1.70E-02</td>
<td>Oct4</td>
<td>3.63E-02</td>
</tr>
<tr>
<td>Smad1</td>
<td>2.79E-02</td>
<td>Zfx</td>
<td>2.96E-02</td>
</tr>
<tr>
<td>Oct4</td>
<td>3.96E-02</td>
<td>Sox2</td>
<td>3.80E-02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Downregulated Genes</th>
<th>24 Hrs ATI</th>
<th>36 Hrs ATI</th>
<th>48 Hrs ATI</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF</td>
<td>p-value</td>
<td>TF</td>
<td>p-value</td>
</tr>
<tr>
<td>Oct4 + Sox2</td>
<td>3.59E-06</td>
<td>Sox2 - Oct4</td>
<td>2.21E-02</td>
</tr>
<tr>
<td>Oct4 + Sox2 + Nanog</td>
<td>6.81E-03</td>
<td>Oct4 + Sox2</td>
<td>2.74E-02</td>
</tr>
<tr>
<td>Smad1</td>
<td>2.68E-02</td>
<td>E2f1</td>
<td>3.45E-02</td>
</tr>
<tr>
<td>Oct4</td>
<td>3.43E-02</td>
<td>n-Myc</td>
<td>4.06E-02</td>
</tr>
</tbody>
</table>

Figure 9. *c-Myc, Smad1 and Oct4 targets are up-regulated while targets co-occupied by Oct4 and Sox2 are down-regulated coincident with Oct4 depletion.* All litters (A-C) were induced with tamoxifen at E6.875. A Compared to phenotypically WT Oct4^WT^ littermates, Oct4 transcript is depleted in Oct4^f/f^;CreER^T2+/-^ embryos. Relative abundance was measured using quantitative-PCR ± s.e.m. B,C After Oct4 depletion up-regulated genes are consistently enriched for targets of c-Myc, Smad1 and Oct4 while down-regulated genes are enriched for targets bound by both Oct4 and Sox2. Enrichment for binding of Nanog, Oct4, STAT3, Smad1, Sox2, Zfx, c-Myc, n-Myc, Klf4, Esrrb, Tcfcp211, E2f1, CTCF, p300, Suz12, Oct4 and Sox2, Oct4 and Nanog, Oct4 and Sox2 and Nanog, Oct4 not Sox2, Oct4 not Nanog, Oct4 not Sox2 not Nanog, Oct4 not c-Myc, Oct4 not Smad1, Sox2 not Oct4, Sox2 not Nanog, Sox2 not Oct4 not Nanog, Nanog not Sox2, Nanog not Oct4, Nanog not Sox2 not Oct4, c-Myc not Oct4 and Smad1 not Oct4 were assessed. B TF binding enrichment amongst up-regulated genes using hypergeometrics. C TF binding enrichment amongst down-regulated genes using hypergeometrics.
of these proteins physically interact with Oct4. Indeed, there was an enrichment for physical interactors amongst genes up-regulated after Oct4 depletion (p=1.99E-08 24hr ATI, p=1.64E-05 36hr ATI, p=5.55E-07 48hr ATI enrichment using hypergeometrics), suggesting that after E7.5 Oct4 represses expression of cofactors it has in ES cells.

To determine whether the differential expression was a direct consequence of Oct4 depletion at its genomic targets, we assessed whether Oct4’s direct targets were enriched amongst up- or down-regulated genes. Systematic mapping of TF targets in early embryos is currently prohibitive [296], so the most extensive existing data set of genome-wide binding maps involving Oct4 in ES cells was used [126]. Enrichment of TF binding targets from ES cells amongst differentially expressed genes after E7.5 requires that binding sites be conserved between these stages. Oct4 binding sites from ES cells were enriched amongst up-regulated genes (Fig. 9B), supporting conservation of the binding sites between ES and ~E8.0 embryos. We suggest that the expression changes at these sites were a direct consequence of Oct4-mediated transcriptional repression being relieved after E7.5.

To test whether signalling networks other than the genomic target set directly mediated by Oct4 might contribute to the Oct4COND MUT phenotype, we determined the transcriptional response that target sets bound by TFs other than Oct4 had to Oct4 depletion. The binding maps of 12 other TFs were assessed for enrichment amongst the genes differentially expressed after Oct4 depletion (see methods for TFs and combinations tested) [126]. Targets of c-Myc and Smad1 were enriched amongst genes up-regulated after Oct4 depletion [126]. Unlike c-Myc which does not cluster at binding sites with Oct4 in the genome, Oct4 facilitates the binding of Smad1 such that they overlap at a subset of sites [126]. However upregulation of Smad1 targets after Oct4 depletion occurred at sites Smad1 occupies independent of Oct4, indicating that
enrichment of up-regulated Smad1 targets is not due to direct relief of Oct4-mediated repression at sites which the two co-occupy [126]. The enrichment of Smad1 targets amongst up-regulated genes that are not co-occupied by Oct4 are: p=6.14E-06 24hr ATI, p=4.55E-03 36hr ATI, p=3.53E-09 48hr ATI (hypergeometric test). These data suggest that the absence of Oct4 yields a transcriptional environment conducive to target activation by c-Myc and Smad1. Conversely, enrichment of sites targeted by both Oct4 and Sox2 amongst down-regulated genes (Fig. 9C) suggests that Oct4 participates in transcriptional activation at these sites after E7.5. Since conditional removal of Sox2 and Pou5f1 do not phenocopy (compare Fig. 3C to 8A), Sox2 is either not essential for activation of these sites, which is consistent with data from ES cells [124], or down-regulation of these targets does not contribute to the Oct4COND MUT phenotype (Fig. 9).

Decreased ‘TGF-beta signalling’ and increased ‘nuclear import of NF-κB’ are primary transcriptional responses of Oct4 depletion.

Oct4 binds thousands of sites in the genome, and it is unlikely that disruption of a single target gene causes the Oct4COND MUT phenotype. To relate molecular changes resulting from Oct4 depletion with the Oct4COND MUT phenotype, we determined which signalling pathways were disrupted coincident with Oct4 depletion and prior to the onset of the phenotype. Unsupervised clustering was used to assess the function of differentially expressed genes collectively. To discern primary effects of Oct4 depletion, we subsetted for genes that are direct targets of Oct4 based on the ES binding maps [126], clustered these (Fig. 10A), and then compared the clusters to global changes (Fig. 10B). 3 of the 4 pathways showing the strongest enrichment in the set of direct targets also showed significant enrichment in the global set. Coordinate regulation of additional genes that are not targets of Oct4 within the same pathways as those directly regulated
by Oct4 suggests amplification of the direct effects (Fig. 10A,B). Quantitative-PCR on
independent biological samples confirmed a subset of changes from the global expression
analysis (Fig. 10C). Notably, PCP signalling components that were candidates for the extension
deficit did not change. The expression profiling suggested that decreased TGF-beta signalling
and increased nuclear import of NF-κB were primary effects as they occurred ~4 hours after
Oct4 depletion amongst the direct targets of Oct4, while decreased Notch signalling and
increased protein translation are other candidates for primary effects that occurred amongst direct
targets of Oct4, albeit 12hrs later (Fig. 10A).
Figure 10. Differential gene expression coincident with loss of Oct4. All litters (A-C) were induced with tamoxifen at E6.875. A Unsupervised clustering of the gene expression changes of Oct4 binding targets, comparing Oct4^{f/f},CreER^{T2+/-} embryos relative to Oct4^{f/f} littermates. Global
effects corresponding to the direct effects at Oct4 binding targets support the utility of sub-setting for Oct4 binding targets in identifying primary effects of Oct4 depletion and the relevance of these primary effects to the Oct4<sup>COND MUT</sup> phenotype in that they appear amplified into effects on the overall gene expression profile (see blue script in panel ‘A’ and ‘B’ for these). The most enriched pathway is provided for each cluster, and an additional pathway provided (in black text) for the cluster where the most enriched pathway did not translate to a global change. B Unsupervised clustering of global expression changes in Oct4<sup>fl/fl;CreER<sup>29+/−</sup></sup> embryos relative to Oct4<sup>fl/fl</sup>. The most enriched pathway and binding factor are provided for each cluster (black text), while primary effects which translated to enriched effects in the global set are in blue text. C Confirmation of expression change for select genes by quantitative PCR in independent litters. Error bars indicate ±s.e.m.

The TGF-β/Nodal signalling pathway maintains embryonic Pou5f1 expression [254, 255], and these data indicate that Oct4 maintains TGF-β signalling which was previously implicated in processes disrupted in Oct4<sup>COND MUT</sup>: appropriate TGF-β signalling is required for the expansion of primitive streak [297], patterning derivatives of the anterior primitive streak [298], establishment of definitive endoderm [299], maturation of the node [300] and left/right asymmetry establishment [301, 302]. TGF-β signalling through Smad2 competes with Smad1 for the co-activator Smad4 [303], so upregulation of Smad1 targets following Oct4 depletion may involve an increase in Smad1, expansion of the domain of activated phosphorylated Smad1 (p-Smad1), or diminished competition from Smad2. Increased abundance of Smad1 was confirmed (Fig. 10C), while the p-Smad1 domain was not altered 24 hours ATI (Fig. 11C,D) or 36 hours ATI (Fig. 12C,D). These data support the increase in Smad1 target activation being within the normal domain present in WT embryos, due to diminished TGF-β signalling that results in decreased competition for Smad4 from Smad2 and upregulation of Smad1.
Figure 11. Decreased proliferation contributes to Oct4COND MUT phenotype (see also Figure 12). Oct4f/f are phenotypically WT controls, and all litters (A–K) induced with tamoxifen at E6.875 and E7.375. A, B Specification of Chordin in the node still occurs after Oct4 depletion (A) Oct4f/f 24 hours ATI. (B) Oct4f/f;CreERT2+/− 24 hours ATI. C, D The expression domain of p-Smad1 is comparable after Oct4 depletion. (C) Oct4f/f 24 hours ATI (D) Oct4f/f;CreERT2+/− 24 hours ATI. E, F Distribution of apoptotic Caspase-3 positive cells 48 hours ATI. (E) Oct4f/f (F) Oct4f/f;CreERT2+/−. G, H Distribution of phospho-histone-3, which marks proliferating cells, 24 hours ATI. Proliferation is significantly reduced in the primitive streak (bracketed by a white line) of Oct4f/f;CreERT2+/− embryos. (G) Oct4f/f (H) Oct4f/f;CreERT2+/−. I Quantification of apoptotic frequency in Oct4f/f and Oct4f/f;CreERT2+/− embryos (F1,44 = 13.16, p<0.05 2-way ANOVA; *p<0.05, **p<0.01 Bonferroni posttest). J Quantification of proliferation frequency in Oct4f/f and Oct4f/f;CreERT2+/− embryos 24 hours ATI (F1,68 = 3.28, p<0.05 2-way ANOVA; **p<0.01 Bonferroni posttest). Oct4 depletion only affects proliferation significantly in the primitive streak. K Distribution of gene expression changes in Oct4f/f;CreERT2+/− embryos. Transcript abundance was quantified 24 hours ATI by quantitative-PCR and relative abundance compared between Oct4f/f;CreERT2+/− and Oct4f/f control embryos (F2,8 = 12.14, p<0.05 2-way ANOVA; ***p<0.001 Bonferroni posttest). All scale bars (A–H) are 50µm and all error bars (I–K) are ±s.e.m.
Current data do not provide a clear role for NF-κB signalling in the processes disrupted in Oct4COND MUT, however there is suggestive evidence: NF-κB binds within the asymmetric enhancer of Nodal, suggesting a role in left-right asymmetry determination [304]; a homologous component of NF-κB in X. laevis is required for mesoderm induction and posterior extension [305]; and the D. rerio homologue coordinates delamination, a process that is critical for mesendodermal movement during gastrulation [306]. A significant decrease in direct targets of Oct4 amongst the Notch pathway occurred later than changes in TGF-β and NF-κB transcript levels. Notch signalling is essential for somitogenesis and its decrease 36 hours ATI coincides with the stage when the first somites should be forming. A decrease in the transcript of Notch signalling component Delta-like 1 (Dll1) was confirmed by QPCR in separate litters (Fig. 10C). Dll1 loss disrupts node formation and is sufficient to cause defects in left/right asymmetry [307]. The node is derived from the anterior of the primitive streak and is required to coordinate left-right asymmetry, specification of definitive endoderm and somitogenesis [308], so its malformation may explain several features of the Oct4COND MUT phenotype, however specification of the node marker Chordin occurs: 24 hours ATI (Fig. 11 A,B) and 36 hours ATI (Fig. 12E,F) [309]. This indicates that at least the earliest stages of node formation occurred normally.
Figure 12. Cdx2, p-Smad1 and Chordin domains are intact in Oct4\textsuperscript{COND MUT}, as is early lineage specification of mesoderm and neuroepithelium. All litters (A-H) were induced with tamoxifen at E6.875 and E7.375. A,B The caudal to rostral gradient of Cdx2 in the posterior persists following Oct4 depletion. (A) Oct4\textsuperscript{f/f} posterior 24hrs ATI. (B) Oct4\textsuperscript{f/f};CreER\textsuperscript{T2+/-} posterior 24hrs ATI. C,D The caudal to rostral gradient of p-Smad1 in the posterior persists following Oct4 depletion. (C) Oct4\textsuperscript{f/f} posterior 36hrs ATI. (D) Oct4\textsuperscript{f/f};CreER\textsuperscript{T2+/-} posterior 36hrs ATI. E,F Specification of Chordin in the node occurs in spite of Oct4 depletion 36hrs ATI. (E) Oct4\textsuperscript{f/f} (F) Oct4\textsuperscript{f/f};CreER\textsuperscript{T2+/-}. G,H Indicators of lineage specification are expressed appropriately following Oct4 depletion. The region outlined with a white box in panel ‘i’ is magnified and provided as an inset in panel ‘ii.’ The heart tube (*) and presumptive forebrain (%) are indicated to provide additional landmarks. G Brachyury is present 48hrs ATI, visible here in trunk mesenchyme (arrow). H Sox2 is present 48hrs ATI throughout neuroepithelium (arrows) and gut (arrowhead). All scale bars (A-H) are 50 µm and WT controls are Oct4\textsuperscript{f/f} in all cases. Sections are oriented with the anterior at the left of each panel.

Oct4 depletion leads to decreased proliferation in the primitive streak

Oct4\textsuperscript{COND MUT} embryos were deficient in primitive streak derived mesenchyme (Fig. 4D), and a general effect on the streak would have broad relevance: cranial mesenchyme supports NTC, while mesendoderm facilitates posterior extension, somitogenesis and turning. The frequency of cells undergoing apoptosis (Caspase-3 positive) in the Oct4\textsuperscript{COND MUT} was increased
(Fig. 11I), and although diminished cell viability might contribute to the phenotype, the random distribution of apoptotic cells throughout the embryo suggests that this was a secondary defect. Conversely, fewer cells proliferated (Phospho-Histone H3 (PH3) positive) in the primitive streak of embryos 24 hours ATI (Fig. 11G,H,J). Reduced proliferation in the primitive streak due to Oct4 removal provides a cellular explanation for failed NTC, posterior extension, turning and somitogenesis. To confirm the localization of these effects, we divided embryos into three segments (proximal anterior, distal and proximal posterior) and quantified the abundance of transcripts regulating apoptosis and proliferation. While there was no difference in the transcript abundance of apoptosis regulators *Bcl-associated X protein* (*Bax*) and *B cell leukemia/lymphoma 2* (*Bcl2*), *Cyclin-dependent kinase inhibitor 1c* (*Cdkn1c*) transcript abundance was selectively increased in the posterior third of embryos coincident with Oct4 depletion (Fig. 11K). These data indicate that ubiquitous Oct4 depletion causes deficient proliferation in the primitive streak.

**Discussion**

After E7.0, Oct4 is still present in the primitive streak, endoderm, several mesoderm derivatives, neuroepithelium as well as extraembryonic endoderm and mesoderm [142]. Derivatives of the primitive streak and developmental processes they regulate are disrupted when Oct4 is depleted ~E7.5: turning, somitogenesis, posterior extension and NTC all are effected, demonstrating for the first time that Oct4 is required in somatic tissue after implantation. The independent effects of Oct4 loss on cranial NTC defect and turning compared to co-occurring effects on caudal NTC and posterior extension might be explained by requirements for Oct4 in different progenitor populations: in definitive endoderm for turning [310], in the node for left-
right asymmetry, in tailbud progenitors for posterior extension and in mesoderm for somitogenesis, posterior extension, left-right asymmetry and bridging cranial neuroepithelium. Each of these progenitor populations is derived from the primitive streak, with the node forming the anterior boundary. Thus a proliferation deficit in the primitive streak which would impair derivatives of all these populations is a sufficient explanation for each aspect of the Oct4\textsuperscript{COND MUT} phenotype, and consistent with the lack of a requirement for Oct4 in extraembryonic tissue. The deficit after E7.5 is not caused by a loss of Oct4 function within the equivalent pluripotency network. Comparable depletion of pluripotency factor Sox2 does not phenocopy Oct4 depletion as it does in pluripotent ES cells. The genomic targets co-occupied by Oct4 and Sox2 in ES cells decrease in expression after Oct4 depletion, which may indicate that Sox2 does not activate these sites after E7.5 or that the Oct4\textsuperscript{COND MUT} phenotype results from its role at sites Oct4 binds independent of Sox2. The primary effects amongst direct genomic targets of Oct4 coincident with its loss are decreased TGF-β signalling and increased signalling promoting nuclear translocation of NF-κB, suggesting candidate mediators of the phenotype. So how do these changes relate to failed expansion of posterior progenitor populations?

Diversion of precursors from one pool to another via cell fate change could cause a progenitor deficiency. Diversion of prospective mesoderm to ectopic neural tubes occurs in several mutants: Fgfr1 [311], Wnt3a [312] and T-box 6 (Tbx6) [313]. Given that Tbx6 represses Sox2 in the tailbud promoting paraxial mesoderm instead of the neural plate [314], that antagonism between Sox2 and Oct4 controls neural versus mesoderm fate during ES differentiation [268], and that XlPou91 (paralogue of Oct4 in X. laevis) promotes mesoderm specification as opposed to a neural fate in response to FGF [265, 315], we followed the fate of putative neuroepithelium and mesoderm after excising Pou5f1. Lineages tracing does not require
complete recombination within a domain, and indeed no lineage divergence was observed when Oct4 was depleted with Sox1-Cre or Bry-Cre. Several pieces of evidence might suggest that cell fate divergence occurs in the conditional approach before recombinase expressed from these lineage-specific loci is active. First, the neuroepithelium of Oct4<sup>COND MUT</sup> is thicker, particularly near closure point 1, although general thickening is common amongst mutants with NTC defects [283] and unrelated to cell fate divergence in such mutants. Second, Sox2 conditional mutants may have the opposite phenotype to Oct4<sup>COND MUT</sup>; in that they appear to have thinner neuroepithelium, but Sox2 is required for maintenance of neural progenitors [316] so this is likely due to progenitor differentiation prior to expansion and unrelated to cellular allocation.

Third, significant alteration in the expression of key determinants of lineage allocation Tbx6, Dll1, Wnt3a and Fgf8 occurs when Oct4 is depleted, however these may just indicate disruptions in the somitogenesis program. The observed proliferation deficit in the primitive streak may mask an effect of Oct4 depletion on segmentation within a given domain, however we saw no evidence of ectopic neural tubes when using CreER<sup>T2</sup>, as would be expected for a diversion of mesoderm to neuroepithelium, nor cell fate switching with the lineage specific recombinases. These data do not exclude an influence of Oct4 on cell fate prior to E7.5, perhaps at the onset of gastrulation (~E6.0) and equivalent to the <i>in vitro</i> effect it has on germ layer specification [268], however, our data do not support the Oct4<sup>COND MUT</sup> phenotype being due to cell fate diversion between germ layers.

In the absence of cell fate diversion, posterior truncation must involve differences in migration, apoptosis and/or proliferation. The distribution of apoptotic cells in Oct4<sup>COND MUT</sup> embryos appears random and is likely a secondary effect. Reduced mesenchyme density, decreased proliferation in the primitive streak and increased <i>Cdkn1c</i> in the Oct4<sup>COND MUT</sup>
embryonic posterior all indicate that expansion of posterior progenitors is disrupted when Oct4 is depleted. One mode by which Oct4 could facilitate progenitor expansion is by promoting the cell cycle, thereby limiting the time required for chromatin remodelling to a lineage-specific program [132, 317]. The G1/S checkpoint is effectively absent from ES cells [132]. Oct4 binds miRNAs which suppress inhibitors of the G1/S transition [132], and in doing so may promote cell cycle progression via bypass of this checkpoint. Also in support of this model, genes regulating ‘chromatin modification’ are up-regulated concurrent with the reduced proliferation that follows Oct4 depletion (cluster 1-295 p=2.1E-04 and cluster 613-908 p=3.7E-04 using hypergeometrics). One point that is not clear is how activation of proliferative agent c-Myc’s targets relates to decreased proliferation: both c-Myc targets co-occupied with Oct4 as well as those bound independently are activated (enrichment amongst upregulated genes is p<5E-03 at all timepoints for both classes using hypergeometrics). c-Myc is part of a module that activates G1/S checkpoint complexes [318, 319]. This module is separate from Oct4. One putative model of why c-Myc target activation increases coincident with the decreased proliferation that accompanies Oct4 depletion is that c-Myc may be required to promote G1/S transitions when the G1/S checkpoint, which is largely absent from undifferentiated populations, is established coincident with Oct4 depletion.

Presence of Pou5f1 in lizards and urodeles is not consistent with its conserved function being in lineage segregation of pluripotent cells from trophectoderm, as neither lizards nor urodeles have trophectoderm [259]. Pou5f1 arose by duplication of Pou2, and the POU-homeodomain and POU-specific domains are found in both [265]. Morrison and Brickman proposed [265] that the evolutionarily conserved role of Oct4 might be facilitating expansion of progenitor populations during and after gastrulation based on work with paralogues: Pou2 in D.
rerio and XlPou91 in X. laevis [265]. Consistent with this, morpholinos against XlPou91 produce a partially conserved phenotype with Oct4COND MUT in that they share posterior truncations and anterior defects [265]. Similarly, embryonic proliferation is impaired in Pou2 mutants prior to gastrulation [261]. We did not explore possible neural segmentation defects in Oct4COND MUT since they have been characterized following disruption of Pou2, XlPou91 and following Oct4 overexpression in mice [260, 265, 320]. XlPou91 loss also promotes endoderm and notochord at the expense of mesoderm [257, 321], but it is not clear whether this is also the case in mice. These data suggest that expansion of all tissues is impacted in Oct4COND MUT, as opposed to diversion from mesoderm to endoderm or a selective effect on certain cell types. Furthermore, these data do not distinguish whether the observed increase in the number of cells in cross-sections of Oct4COND MUT notochords are due to altered cell fate specification or an apparent increase that is actually due to deficient morphogenetic extension. However, since the notochord narrows during extension, and deficient proliferation in the primitive streak would contribute to the observed extension deficit that at least contributes to the broader notochord. Pou2 mutants (D. rerio) also have posterior truncations, although the cell lineage influence is nearly an inversion of that in either mice or X. laevis with its loss promoting mesoderm as opposed to endoderm and its presence being required for endoderm as opposed to mesoderm [263, 322]. Conserved phenotypes between Oct4COND MUT and XlPou91 (X. laevis) mutant embryos parallels rescue experiments where XlPou91 rescues Oct4−/− ES cells, but Pou2 (D. rerio) does not [265]. These data suggest a structural, as opposed to a contextual, difference between Oct4/XlPou91 and Pou2 is responsible for the differences in the loss of function phenotypes. Differences in cell fate influence aside, disruption of each homologue results in reduced proliferation and ultimately differentiation. This supports a conserved role for Oct4 in ensuring sufficient expansion of
progenitor populations – in mice these include undifferentiated preimplantation populations through progenitors in the primitive streak. Indeed, the present study points to a specific proliferative defect in the primitive streak as a single explanation for the multiple embryonic phenotypes seen after *Pou5f1* excision at E7.0.
Chapter 3: Surfaceome profiling reveals novel regulators of neural stem cell function.

I contributed to this project by helping design experiments, collecting the data aside from the mass-spec data (collection, normalization and protein prediction by Damaris Bausch-Fluck) and authoring the manuscript.

This chapter has been submitted for publication:
Abstract

The composition of cell-surface proteins changes during lineage specification, altering cellular responses to their milieu. The changes that characterize maturation of early NSCs remain poorly understood. Here we use mass-spectrometry (MS) based Cell Surface Capture (CSC) technology to profile the cell surface of early NSCs and demonstrate functional requirements for several enriched molecules. Primitive NSCs arise from ES cells upon removal of TGF-β signalling, whilst definitive NSCs arise from primitive NSCs upon Lif removal and FGF addition. In vivo aggregation assays revealed that N-cadherin upregulation is critical for the initial exclusion of definitive NSCs from the pluripotent population, while c-Kit signalling mediates quiescence of primitive NSCs. Further, we implicate Ephrin A4 (EphA4) in NSC survival signalling and Erbb2 as being required for primitive NSC proliferation. This work elucidates several key mediators of NSC function whose relevance is confirmed on forebrain-derived populations and identifies a host of other candidates that may regulate NSCs.

Introduction

Mammalian development employs regulative aspects, whereby exogenous signals instruct multipotent precursors on cell fate decisions. As development progresses lineages mature, acquiring cell-autonomous properties including distinct adhesive profiles that facilitate morphogenetic movements, boundary formation as well as cell-intrinsic lineage programs which restrict the range and output of morphogen responsiveness. Cell surface protein composition at the earliest stages of NSC development is poorly understood, and improved characterization would help elucidate the regulators of critical NSC properties.
The prevailing model of neural induction posits that neural tissue is segregated from ectoderm by release of bone morphogenetic protein (BMP) antagonists such as Noggin and Chordin from anterior visceral endoderm at E7.5 of murine development [323]; BMP antagonism effectively relieves repression of neural fate, as factors downstream of BMPs including SMADs and inhibitors of differentiation (ID) inhibit the neural program [111, 324]. These events in development can be modeled using pluripotent ES cells that self-renew when cultured with BMP4 and Lif [325], but undergo neural induction when placed at low density in serum free media with Lif but without exogenous BMPs [324]. Greater than 90% of ES cells which survive the transition into serum free media express the neural marker Sox1 after 20hrs [326], consistent with Lif signalling through Stat3 mediating a block of non-neural differentiation via Nanog [111, 325, 327, 328].

The earliest NSCs arise at E6.0 of murine development [329]. This Lif-dependent population decreases in abundance between E7.5 and E8.5 of development, coincident with an increase in the prevalence of FGF-dependent NSCs that persist through adulthood [329]. When ES cells are transferred from self-renewal conditions to serum-free media with Lif, a subset referred to as primitive NSCs (pNSCs) proliferate extensively as clones [324]. pNSCs then give rise to definitive NSC (dNSCs) that are not Lif-dependent, but require exogenous FGF comparable to embryo-derived NSCs [324]. FGF is required for proliferation of neural populations [326, 330] and downstream signalling via ERK—at least transiently—contributes to neural maturation [331]. We reasoned that an unbiased profile of the cell-surface constituents of each cell in the early NSC lineage would elucidate the molecules that dictate the identity of each cell type. In future, such profiles could be used to systematically explore the biology of each cell
type and would provide utility in indirect applications such as sorting to enrich or deplete populations of interest.

Here we profile the cell surface of clonally derived populations of ES cells, pNSCs and dNSCs in this sequential neural lineage using CSC technology to identify glycosylated plasma membrane proteins with MS [332]. We also demonstrate the utility of this data in making predictions about protein function in NSCs. Specifically, we demonstrate that Neural-cadherin (N-Cad) is the predominant determinant of dNSC compartmentalization, c-kit signalling promotes pNSC quiescence, signalling through EphA4 is required for viability of NSC populations and that Erbb2 signalling is required for the proliferation of pNSCs. Moreover, we show the relevance of these findings to NSCs derived from adult mice.

**Materials and Methods**

**Cell Culture**

R1 ES cells (a gift from A. Nagy) were used for experiments unless otherwise specified. YC5/eYFP ES cells (a gift from A. Nagy) were used to generate pNSC and dNSCs for aggregation experiments [333]. For maintenance, ES cells were grown on mitotically inactive embryonic fibroblast layers in DMEM plus 15% fetal calf serum (FCS)(Hyclone) culture medium containing LIF (1000 U/ml). pNSCs and dNSCs were generated as described in (Tropepe et al., 2001).

**Inhibitor Assays**

88
Inhibitors were added to the cultures during plating. N-cad function blocking antibody (Sigma), EphA4 blocking peptide (Alta Biosciences), Erbb2 inhibitor II (Calbiochem), Gleevec (Toronto Research Chemicals Inc.) or Cyclopamine (Toronto Research Chemicals Inc.) that were dissolved in PBS, PBS, DMSO, DMSO and EtOH respectively. Viability was assessed using trypan blue exclusion after 24hrs. Proliferation deficits were inferred in situations where inhibitors did not cause significant effects on viability but did prevent sphere formation. Viability and proliferation of ES cells was assessed on both gelatin (Millipore) and mitotically inactivated feeders.

Cell surface capturing

Cell surface capturing was performed as described previously [332]. In brief, colonies were dissociated by manual trituration after incubation in 1mM EDTA and 10% FBS (Hyclone) for 15min at 37°C. Extracellularly exposed aldehydes were then oxidized with 1.6mM sodium metaperiodate (NaIO4) (Piercenet) and reacted for 1 hour with 5mM biocytin hydrazide (Biotium.com). The cells were then washed and lysed by dounce homogenization in a hypotonic lysis buffer and the nuclei were pelleted by centrifugation. The supernatant containing the membranes and the cytoplasm was mixed with membrane prep buffer (280 mM sucrose, 50 mM MES pH 6, 450 mM NaCl, 10 mM MgCl2) and subjected to ultra centrifugation. The microsomal pellet was collected after 1 hour centrifugation at 35 000 rpm and solubilized by addition of 0.1% RapiGest (Waters) and sonification. After overnight trypsin digestion, the biontynlated peptides were coupled to streptavidin beads (Piercenet), thoroughly washed and released by PNGaseF (NEB), which cleaves N-glycosylated peptides at their backbone. Peptides
were then clean-up over C18-tips (The NestGroup) and subjected to reverse-phase liquid chromatography coupled to tandem mass-spectrometry.

**Mass spectrometry**

All MS data from this study can be downloaded in the open source mzXML format from (http://www.peptideatlas.org/repository/) and was integrated into the publicly accessible PeptideAtlas database at (http://www.peptideatlas.org/).

Each peptide sample was analyzed on an Eksigent Nano LC system (Eksigent Technologies) connected to a hybrid linear ion trap LTQ Orbitrap (Thermo Scientific) mass spectrometer, which was equipped with a nanoelectrospray ion source (Thermo Scientific). Peptide separation was carried out on a RP-HPLC column (75 µm x 10 cm) packed in-house with C18 resin (Magic C18 AQ 3 µm, Michrom BioResources) using a linear gradient from 90 % solvent A (water, 0.2 % formic acid, 1 % acetonitrile) and 10 % solvent B (water, 0.2 % formic acid, 80 % acetonitrile) to 65 % solvent A and 35 % solvent B over 60 min at a flow rate of 0.2 µl/ min. The data acquisition mode was set to acquire one high resolution MS scan in the ICR cell followed by three collision induced dissociation MS/MS scans in the linear ion trap. For a high resolution MS scan, $10^6$ ions were accumulated over a maximum time of 500 ms and the FWHM resolution was set to 60,000 (at m/z 300). Only MS signals exceeding 500 ion counts triggered a MS/MS attempt and $10^4$ ions were acquired for a MS/MS scan over a maximum time of 250 ms. The normalized collision energy was set to 35% and one microscan was acquired for each spectrum. Singly charged ions were excluded from triggering MS/MS scans.
All acquired MS/MS spectra were searched against the International Protein Index (IPI) database (Version 3.26) using the Sequest algorithm. The Sequest database search criteria included: 0.2 Da mass tolerance for the precursor ion, 0.5 Da mass tolerance for the fragment ions, variable modifications of 0.984016 Da for asparagines (representing formerly N-glycosylated asparagines after deamidation through the PNGaseF treatment) and 15.994915 Da for methionines (covering rapidly oxidizing methionines), carbamidomethylation as static modification for cysteines, at least one tryptic terminus, two missed cleavage sites. Statistical analysis of the data were performed using the Trans-Proteomic Pipeline v4.3 (TPP) including PeptideProphet and ProteinProphet. The ProteinProphet probability score was set as such, that the false discovery rate was less than 1% determined by ProteinProphet. All MS data from this study can be downloaded in the open source mzXML format from (http://www.peptideatlas.org/repository/) and was integrated into the publicly accessible PeptideAtlas database at (http://www.peptideatlas.org/).

**Label-free Quantification**

Protein quantification was performed using Progenesis (Nonlinear Dynamics). After manually improving the alignment, quantified peaks were filtered for identification by sequence search and overall protein abundances were calculated thereof. ANOVA tests were applied to determine the significance of the detected differences.

**Animal Husbandry**

All of the mice used in this study were CD1 mice obtained from the Jackson laboratories and housed in accordance with the Institutional and Governmental Animal Care Committee.
guidelines of the University of Toronto.

**Morula Aggregations**

8-cell ES, pNSC or dNSC colonies were aggregated with diploid CD1 morula for 24 hr (Nagy and Rossant, 2000). Embryos were then transplanted into pseudo-pregnant CD1 females, and removed at 9.5 dpc or taken as postnatal animals and assessed for fluorescent contribution by fluorescent microscopy or as described in ‘Quantification of Chimerism.’

**N-Cadherin Aggregation Assay**

pNSCs were passaged into SFM +FGF2 +B27 as described previously, with the addition of 0.8ug/ml N-Cad function-blocking antibody (Sigma) during plating. 0.8ug/ml was also added to KSOM during the overnight aggregation.

**Imaging**

Sections of chimeras were imaged using a 40x/0.60 Olympus IX81 inverted microscope with the Olympus Microsuite Version 3.2 Analysis imaging system software (Soft Imaging System). Overnight integration into the ICM was visualized with the same physical setup and Olympus FV10-ASW v2.01 software, and confirmed in a subset of cases with confocal imaging. Sections of embryoid bodies were taken with Zeiss Axio Observer.D1 inverted fluorescent microscope equipped with an AxioCamMrm digital camera run by AxioVision v4.8 software.

**Confocal**

Images of dsRed chimeric embryos were taken every 5 µm with a Leica TCS2 confocal
microscope.

**Immunocytochemistry**
Sections were permeabilized with 0.3% Triton X-100 detergent (Sigma), sequentially blocked with 1% bovine serum albumin (Sigma), followed by 10% normal goat serum (NGS)(Sigma) in Stockholm's PBS + 0.3% Triton X, pH 7.3 for 30 min at room temperature before application of primary antibodies overnight in Stockholm's PBS, 1.0% NGS, and 0.3% Triton X-100. Primary anti-smooth muscle actin was applied at 1:250 (Sigma). Sections were washed and re-blocked as above before applying the secondary antibodies at 37°C for 2 hr in Stockholm's PBS + 1.0% NGS. Goat anti-mouse 568 nm Alexa Fluor antibodies (1:333; Invitrogen) was used as a secondary. Nuclei were counterstained with 10 µg/ml Hoechst 33258 (Sigma) before samples were mounted and coverslips applied using Gel Mount (Biomed a).

**Quantification of Chimerism**
To determine the extent of chimerism, animals were either sacrificed for aNSCs isolation or perfused with 2% paraformaldehyde (Sigma). Tissues of interest were then rinsed, equilibrated in 30% sucrose, mounted in cryoprotectant and sections of 12 µm taken on a Jencon's OTF5000 cryostat.

Chimerism in the cortex, heart and epithelium of the gut was assessed by counting the fraction of nuclei that were eYFP+. Chimerism within smooth muscle of the gut was taken to be the fraction of eYFP+ area within the total smooth muscle actin positive area, measured using ImageJ.
Finally, chimerism within the SE of the lateral ventricles was scored as the fraction of eYFP+ spheres.

**Statistics**

ANOVA were used to assess the significance of inhibitor effects on each cell type (two-way with Bonferroni), differences in differentiation potential between cells in the lineage (two-way with Bonferroni) as well as the difference in abundance of proteins between cell types (one-way for each protein with Bonferroni). A significance threshold of p<0.05 was used for all tests.

**Results**

**Cell surface profiling reveals distinct NSC signatures**

Transcriptional profiling is not an ideal predictor of protein profiles, evidenced by significant discrepancies between levels of protein and corresponding mRNAs [334]. These data suggest that meaningful regulation of protein profiles occurs at the level of translation and supports the merit of directly assessing proteins for both identification and quantification.

During development, lineage maturation and terminal differentiation are governed by a variety of factors including adhesion molecules that dictate compartmentalization and receptors that dictate morphogen responsiveness. Proteins at the cell surface are of particular interest in understanding these processes since they mediate cell-cell communication, signal transduction, migration and adhesion. To elucidate plasma membrane proteins involved in maturation of the neural lineage, we compared the cell surface protein profile of ES, pNSCs and dNSCs using CSC [332, 335]. CSC offers a major advantage over antibody profiling approaches to cell-surface
protein characterization in that CSC technology enables protein identification and relative quantification of the broad spectrum of plasma membrane glycoproteins in parallel.

Unlike the inner side of the plasma membrane which lacks conventional-type oligosaccharides, the external surface is rich in carbohydrates [336]. Most proteins --91.7% (confirmed) - 94.0% (candidate)-- presented at the external surface of mammalian cells contain carbohydrates [337], with the majority of these (97.5%) being N-glycosylated (asparagine-linked oligosaccharides present at the peptide sequence ‘NxS/T’ where x is any amino aside aside from proline). Glycosylation is believed to aid protein folding/assembly within the endoplasmic reticulum, migration to the cell surface, and provide specificity to structures recognized by external ligands [337].

As such, CSC offers insight into the complement of environmental stimuli to which cells may be responsive independent of the antibodies available for detection. Moreover, the technique distinguishes itself from comparable MS-based techniques in that labeling is performed on viable cells. In brief, the glycans of plasma membrane proteins were oxizided with sodium metaperiodate (NaIO4) and labeled with biocytin hydrazide, glycopeptides were purified with streptavidin, then enzymatically released and identified by reverse-phase liquid chromatography coupled to tandem MS (LC MS/MS).

Using CSC and LC MS/MS, 813, 1134 and 880 peptides were detected in the three respective cellular subsets (ES, pNSCs, dNSCs). Peptide and Protein identities were scored using PeptideProphet and ProteinProphet as part of the Trans-Proteomic Pipeline v4.3 (TPP) pipeline [338-340]. Because many peptides map to the same protein, a total of 378, 563 and 456 proteins were identified using Occam’s razor approach of ProteinProphet with a false-discovery rate
(FDR) of <0.01 (Fig 1A). In each cell type over 200 membrane glycoproteins were identified, many of these are shared between the three cell types (Fig 1B).

In order to quantify the differential expression of the cell surface proteins, we analyzed them by label-free quantification [341] with Progenesis software. We focused on those proteins that were identified via at least one peptide containing the N-glycosylation motif, asparagine, any amino acid, serine or threonine (NXS/T) and a deamidation on this asparagine. A distinguishing feature of cell surface proteins compared to cytoplasmic proteins is the fact that most of them are N-glycosylated [342, 343]. Since we specifically target, enrich and modify such glycopeptides during the CSC workflow, their specific and repetitive identification is strong evidence of the cell surface location of the respective protein at the time of labeling.

Overall, 228 proteins were quantified which differed significantly (p<0.05) in relative abundance between at least two of the three cell types. Use of precursors from within the same sequential lineage as background during comparisons with NSCs should enrich for differences
Figure 1. Overview of CSC-based protein identifications. A Summary of the experimental replicates, as well as aggregate peptide and protein identifications in each cell type. B Schematic depicting the number of unique membrane proteins, predicted by Transmembrane Hidden Markov Model, identified in each cell type by ProteinProphet and the overlap between them. C The schematic depicts proteins (grey balls) which differ significantly in abundance between at
least two cell types in the lineage (p<0.05), and are 5 times more abundant in the cell type (blue
balls) to which they are connected with an edge than those to which they are not connected.

that define the biology of each distinct NSC in the lineage. Proteins that are significantly
(p<0.05) different and a minimum of 5 times more abundant based on the MS signal in one cell
type in the lineage than in either of the other two are depicted in Figure 1C (150 proteins).
Proteins are connected by an edge to the cell type in which they are 5 times higher [344].
Similarly, amongst proteins detected and quantified at a confidence score of >1.3 using the TPP
in combination with Progenesis software, the top 20 proteins by fold enrichment in NSCs are
summarized in Table 1-3.

Cell surface protein profiling yielded enrichment of several proteins already implicated in
NSC biology. CD24, a marker of type A neuroblasts that is expressed specifically in the adult
subependymal zone [345, 346], was 25 times higher (p<0.05) in dNSCs compared to ES
populations based on the MS signals, indicating a relative upregulation of the protein. While the
ratio of signal intensity between cell types is not an exact readout of the fold difference in copy
number for a particular protein between cell types, it is an accurate predictor of up- or down-
regulation and can be used to estimate the magnitude of these differences. Prominin, a cell
surface marker commonly used for neural precursor enrichment [347] exhibited 50 times the MS
signal intensity in pNSC than ES cells and 95 times the signal intensity in dNSC than ES cells.
Additionally, several components of the hedgehog signalling pathway, known to be important in
NSC function, were upregulated in NSCs; these included Hedgehog-interacting protein (HhIP)
which is 813 times higher in dNSC than ES cells (p<0.05), as well as Smoothened (Smo) which
is 18 times higher in dNSCs than ES cells and Patched 1 which is 12 times higher in dNSCs than
ES cells. While the numerical accuracy of differences in protein abundance between cell types would be improved by additional repetitions and absolute

<table>
<thead>
<tr>
<th>Rank</th>
<th>Gene Symbol</th>
<th>David Gene Name</th>
<th>Protein ID</th>
<th>Peptide Count</th>
<th>Relative Abundance: pNSC/ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lepre1</td>
<td>leprecan 1</td>
<td>0.9943</td>
<td>1</td>
<td>7528.16</td>
</tr>
<tr>
<td>2</td>
<td>Sk04a1</td>
<td>solute carrier organic anion transporter family, member 4a1</td>
<td>0.9949</td>
<td>1</td>
<td>1000.41</td>
</tr>
<tr>
<td>3</td>
<td>Eda2r</td>
<td>ectodysplasin A2 isoform receptor</td>
<td>1</td>
<td>1</td>
<td>604.58</td>
</tr>
<tr>
<td>4</td>
<td>Sk24a6</td>
<td>solute carrier family 24 (sodium/potassium/calcium exchanger), member 6</td>
<td>0.9961</td>
<td>1</td>
<td>604.48</td>
</tr>
<tr>
<td>5</td>
<td>Alg9a</td>
<td>autophagy-related 9A (yeast)</td>
<td>0.9967</td>
<td>1</td>
<td>488.26</td>
</tr>
<tr>
<td>6</td>
<td>Rnfg</td>
<td>RNF51 O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase; similar to radical fringe</td>
<td>0.996</td>
<td>1</td>
<td>400.76</td>
</tr>
</tbody>
</table>

Table 1. Cell surface proteins showing the greatest fold enrichment in pNSCs relative to ES cells amongst those with a Protein probability of >0.8 in Proteinprophet and meeting a significance threshold of p≤0.05 in a one-way ANOVA with Bonferroni correction assessing the difference in quantity.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Gene Symbol</th>
<th>David Gene Name</th>
<th>Protein ID</th>
<th>Peptide Count</th>
<th>Relative Abundance: dNSC/ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hhip</td>
<td>Hedgehog-interacting protein</td>
<td>0.9966</td>
<td>1</td>
<td>812.95</td>
</tr>
<tr>
<td>2</td>
<td>LOC100045658</td>
<td>CAS1 domain containing 1; similar to O-acetyltransferase</td>
<td>0.9967</td>
<td>1</td>
<td>221.41</td>
</tr>
<tr>
<td>3</td>
<td>Sk22a23</td>
<td>solute carrier family 22, member 23</td>
<td>1</td>
<td>1</td>
<td>167.56</td>
</tr>
<tr>
<td>4</td>
<td>Tmem106b</td>
<td>transmembrane protein 106b</td>
<td>0.9967</td>
<td>1</td>
<td>126.30</td>
</tr>
<tr>
<td>5</td>
<td>Fat1</td>
<td>follistatin-like 1</td>
<td>0.9943</td>
<td>1</td>
<td>99.77</td>
</tr>
<tr>
<td>6</td>
<td>Pxdn</td>
<td>peptidase homolog (Drosophila)</td>
<td>1</td>
<td>1</td>
<td>86.86</td>
</tr>
<tr>
<td>7</td>
<td>Acvr2a</td>
<td>activin receptor IA</td>
<td>1</td>
<td>1</td>
<td>76.76</td>
</tr>
<tr>
<td>8</td>
<td>Col2a1</td>
<td>collagen, type II, alpha 1</td>
<td>0.9962</td>
<td>1</td>
<td>71.04</td>
</tr>
<tr>
<td>9</td>
<td>Cdhh2</td>
<td>cadherin 2; N-cadherin</td>
<td>0.9966</td>
<td>1</td>
<td>52.67</td>
</tr>
<tr>
<td>10</td>
<td>Hgsnat</td>
<td>heparan-alpha-glycosaminide N-acetyltransferase</td>
<td>0.9967</td>
<td>1</td>
<td>38.67</td>
</tr>
<tr>
<td>11</td>
<td>Frem2</td>
<td>Fras1 related extracellular matrix protein 2</td>
<td>1</td>
<td>1</td>
<td>48.00</td>
</tr>
<tr>
<td>12</td>
<td>Ckn5</td>
<td>chloride channel 5</td>
<td>1</td>
<td>2</td>
<td>42.99</td>
</tr>
<tr>
<td>13</td>
<td>Tmed4</td>
<td>transmembrane emp24 protein transport domain containing 4</td>
<td>1</td>
<td>1</td>
<td>40.88</td>
</tr>
<tr>
<td>14</td>
<td>Sema4c</td>
<td>semaphorin 4C; isoform 3</td>
<td>0.9963</td>
<td>2</td>
<td>39.09</td>
</tr>
<tr>
<td>15</td>
<td>Tsapn13</td>
<td>tetraspan 13</td>
<td>1</td>
<td>2</td>
<td>38.79</td>
</tr>
<tr>
<td>16</td>
<td>Dgcr2</td>
<td>DGCR2/George syndrome critical region gene 2</td>
<td>0.9967</td>
<td>1</td>
<td>36.50</td>
</tr>
<tr>
<td>17</td>
<td>Cd24a</td>
<td>CD24a antigen</td>
<td>0.9997</td>
<td>4</td>
<td>25.17</td>
</tr>
<tr>
<td>18</td>
<td>Sk38a2</td>
<td>solute carrier family 38, member 2</td>
<td>1</td>
<td>5</td>
<td>23.70</td>
</tr>
<tr>
<td>19</td>
<td>Tmem181a</td>
<td>predicted gene, 547127; transmembrane protein 181</td>
<td>1</td>
<td>2</td>
<td>22.60</td>
</tr>
<tr>
<td>20</td>
<td>Erbb2</td>
<td>erb-b2 erythroblastosis leukemia viral oncogene homolog 2; neuro/glioblastoma derived oncogene homolog (avian)</td>
<td>0.9961</td>
<td>1</td>
<td>20.48</td>
</tr>
</tbody>
</table>
Table 2. Cell surface proteins showing the greatest fold enrichment in dNSCs relative to ES cells amongst those with a Protein probability of >0.8 in Proteinprophet and meeting a significance threshold of p≤0.05 in a one-way ANOVA with Bonferoni correction assessing the difference in quantity.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Gene Symbol</th>
<th>David Gene Name</th>
<th>Protein ID Probability</th>
<th>Peptide Count</th>
<th>Relative Abundance: dNSC/pNSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cdh2</td>
<td>cadherin 2; N-cadherin</td>
<td>0.9966</td>
<td>1</td>
<td>5850.81</td>
</tr>
<tr>
<td>2</td>
<td>Foxn3</td>
<td>forkhead box N3</td>
<td>0.9384</td>
<td>1</td>
<td>2159.07</td>
</tr>
<tr>
<td>3</td>
<td>Calr</td>
<td>calcitonin receptor-like</td>
<td>0.9957</td>
<td>1</td>
<td>983.60</td>
</tr>
<tr>
<td>4</td>
<td>Slc22a23</td>
<td>solute carrier family 22, member 23</td>
<td>1</td>
<td>1</td>
<td>557.55</td>
</tr>
<tr>
<td>5</td>
<td>Slc1a2</td>
<td>solute carrier family 1 (glial high affinity glutamate transporter), member 2</td>
<td>0.9966</td>
<td>1</td>
<td>84.45</td>
</tr>
<tr>
<td>6</td>
<td>Lpar6</td>
<td>purinergic receptor P2Y, G-protein coupled, 5</td>
<td>0.9966</td>
<td>1</td>
<td>78.94</td>
</tr>
<tr>
<td>7</td>
<td>Erbb2</td>
<td>v-erb-b2 erythroblast viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)</td>
<td>0.9961</td>
<td>1</td>
<td>65.59</td>
</tr>
<tr>
<td>8</td>
<td>Colec12</td>
<td>collectin sub-family member 12</td>
<td>1</td>
<td>3</td>
<td>57.94</td>
</tr>
<tr>
<td>9</td>
<td>Urm1</td>
<td>leucine rich repeat protein 1, neuronal</td>
<td>0.9959</td>
<td>1</td>
<td>52.00</td>
</tr>
<tr>
<td>10</td>
<td>Bb3d17</td>
<td>BTB (POZ) domain containing 17</td>
<td>1</td>
<td>1</td>
<td>51.02</td>
</tr>
<tr>
<td>11</td>
<td>Sort1</td>
<td>sortin 1</td>
<td>0.9702</td>
<td>1</td>
<td>24.37</td>
</tr>
<tr>
<td>12</td>
<td>Gfralpha3</td>
<td>glial cell line derived neurotrophic factor family receptor alpha 1</td>
<td>0.9966</td>
<td>1</td>
<td>20.91</td>
</tr>
<tr>
<td>13</td>
<td>Hsgnat</td>
<td>heparan-alpha-glicosaminidase N-acetyltransferase</td>
<td>0.9967</td>
<td>1</td>
<td>20.48</td>
</tr>
<tr>
<td>14</td>
<td>Sema4c</td>
<td>semaphorin 4C, isoform 3</td>
<td>0.9963</td>
<td>2</td>
<td>18.80</td>
</tr>
<tr>
<td>15</td>
<td>Hhip</td>
<td>Hedgehog-interacting protein</td>
<td>0.9966</td>
<td>1</td>
<td>18.01</td>
</tr>
<tr>
<td>16</td>
<td>Tmed4</td>
<td>transmembrane emp24 protein transport domain containing 4</td>
<td>1</td>
<td>1</td>
<td>9.83</td>
</tr>
<tr>
<td>17</td>
<td>Lrrn4</td>
<td>leucine rich repeat neuronal 4</td>
<td>0.9967</td>
<td>1</td>
<td>8.98</td>
</tr>
<tr>
<td>18</td>
<td>Uman2</td>
<td>actin, mannose-binding 2</td>
<td>0.9937</td>
<td>1</td>
<td>8.57</td>
</tr>
<tr>
<td>19</td>
<td>Pxdn</td>
<td>peroxidasin homolog (Drosophila)</td>
<td>1</td>
<td>4</td>
<td>8.15</td>
</tr>
<tr>
<td>20</td>
<td>Tor1ap1</td>
<td>torsin A interacting protein 1</td>
<td>0.8716</td>
<td>1</td>
<td>7.28</td>
</tr>
</tbody>
</table>

Table 3. Cell surface proteins showing the greatest fold enrichment in dNSCs relative to pNSCs amongst those with a Protein probability of >0.8 in Proteinprophet and meeting a significance threshold of p≤0.05 in a one-way ANOVA with Bonferoni correction assessing the difference in quantity.

quantitative Selective Reaction Monitoring assays using internal standards [348] they are useful estimates and suggest that the profiling data are likely to contain other molecules relevant to the biology of NSCs.

N-Cadherin mediates dNSC compartmentalization

The greatest relative difference in abundance detected between dNSCs and their precursor pNSCs through CSC profiling was in N-cad, a calcium dependent adhesion molecule where we detected an MS signal 5850 times higher in dNSCs (Table 3). Adhesive differences
between cell types are essential during development to organize tissues and establish signalling environments or niches. To assess whether the differences in adhesion molecules detected by mass spectrometry are functionally relevant and constitute candidate mediators of early neural organization, we performed morula aggregation assays. Knowing that ES cells will colonize the ICM when aggregated with morula, we asked whether the other two neural populations in the lineage were equally proficient in doing so.

pNSC colonies were comparable to ES cells, incorporating within the ICM 88% (n=1302/1486) of the time compared to 89% for ES (n=146/164) (Fig 2B). pNSCs aggregation in the ICM enabled assessment of their lineage differentiation potential. Despite pNSCs resolving
Figure 2. NSCs lose the ability to colonize the ICM as they mature through the lineage. 
A Contribution of YFP+ ES cells (green) to the ICM (*) is apparent in 2 of 3 blastocysts in this representative image. B Quantification of the frequency with which donor cells colonize the embryo at the developmental timepoints indicated. Contribution at E3.5 was scored positive if integration within the ICM occurred after overnight integration, at E9.5 if embryos were partially fluorescent and postnatally by pigmented contribution to the coat. Note that embryos were sorted at E3.5 such that only embryos with donor contribution to the ICM were reimplanted into surrogate mothers. The latter two timepoints reflect this. pNSCs persisted to E9.5 in 48 of 52 cases compared to 9 of 19 blastocysts colonized by NSC +Lif +Fgf, and 1 of 5 in the case of dNSCs. 5 of 15 pNSC-colonized blastocysts had chimeric contribution at birth, whereas ES cells persisted to birth in 7 of 16 cases. No postnatal data were available for dNSCs on account of their infrequent colonization of the ICM. C The relative abundance of E-cad was measured using quantitative PCR (± s.e.m.). The data are normalized to abundance in ES cells.

The relative abundance of E-cadherin (E-cad) was measured using quantitative PCR (± s.e.m.). The data are normalized to abundance in ES cells.

To assess whether the difference in adhesive compatibility between pNSC and dNSC was due to the loss of Lif signalling or addition of FGF + B27, we assessed the adhesive properties of pNSCs passaged into Lif, FGF and B27. Under these culture conditions, NSCs colonized the ICM in 51% (n=391/769) of cases indicating that loss of Lif-based signalling is an important, though not the sole, determinant in the altered adhesive profile of dNSCs (Fig 2B).

E-cadherin (E-cad) would be a candidate mediator of this Lif-mediated rescue if dNSCs are excluded from the ICM based on lack of an adhesion factor, as it is primarily responsible for
adhesion of preimplantation embryos [349] and E-cad transcript abundance decreases as the NSC lineage matures in vivo [350]. One limitation of the glycoprotein focused CSC assay was that we were unable to quantify differences in E-cad on account of the NXS/T motif used to select membrane proteins being absent. This is true for a small subset of surface bound proteins. Given the absence of an NXS/T motif in E-cad, we estimated its relative abundance between cell types in the NSC lineage by quantitative PCR (Fig 2C). Lif addition to dNSC culture conditions did not maintain E-cad levels comparable to those in ES and pNSCs (p<0.05). In fact, there was no difference in the levels of E-cad when comparing Lif addition to dNSC culture and dNSCs, indicating that Lif does not rescue the ability of dNSCs to colonize the ICM by maintaining E-cad levels.

To test the prediction that adhesion molecules showing elevated expression in dNSCs relative to ES and pNSCs mediated the exclusion of dNSCs from ICM integration, we blocked N-cad, which showed the greatest difference in abundance between dNSCs and their precursor population, pNSCs. We previously observed upregulation of N-cad in embryonic and adult NSC populations relative to ES cells suggesting that this feature of the in vitro model recapitulates development [350]. Culturing dNSCs in an N-cad function-blocking antibody partially ‘rescued’ their ability to colonize the ICM (Fig 5B-D). Confirmation that rescued cells made bona fide contributions to the ICM was achieved by confocal microscopy in all 8 samples that were optically sectioned. A sample confocal stack of rescued dNSC aggregation is shown in
Figure 3. Lineage distribution bias of pNSC in chimeras. pNSCs persist in 33% (n=5/15) of viable postnatal animals, comparable to the 44% (n = 7/16) observed for ES cells. To assess whether there was an effect on cell fate, we quantified the contribution of chimeric pNSCs and ES cells to a panel of tissues in postnatal animals. Distribution bias when comparing the two populations would indicate a relative difference between ES and pNSCs in tissue allocation or maintenance therein. A, B, C, D Sample images of pNSC (yellow) contribution to postnatal chimeras. A Cerebral cortex B Heart C Small Intestine D SE-derived neurospheres. E, F Quantification of ES and pNSC contribution within a given cohort. E Cohort 1 F Cohort 2.
Figure 4. pNSC distribution is not the result of fusion. To assay whether fusion between donor and endogenous cells was occurring at a high enough frequency to confound our observations, we aggregated dsRED embryos with GFP+ donor cells, but did not see any evidence of fusion. Two days after aggregation, 18 blastocysts with pNSC contribution were scored using optical sections taken on a confocal microscope. We found mutually exclusive expression of dsRED and GFP, with no instances of yellow cells indicating that the frequency of early fusion events is too rare to mask lineage bias or confound our interpretations. A representative section is displayed.

Fig 5D. This result suggests that upregulation of N-cad is an important determinant of early neural progression to a fate that is incompatible with blastocyst integration.

Cell surface protein profiling identifies molecules required specifically for NSC function

We reasoned that molecules enriched in NSCs relative to the ES cells from which they were derived might mediate functions specifically in NSCs. For further analysis, we selected candidates that are 5 times enriched in either pNSCs or dNSCs relative to ES cells (p<0.05) and
Figure 5. Upregulation of N-cad prevents dNSCs from aggregating with morula A 3-dimensional peaks used to quantify the abundance of N-cad. The relative MS integrated peak area signal of the N-cad peptide in the ES cell sample (left) is compared to the equivalent peak area signal in the dNSC sample (right). The peak area within the red box was selected, fragmented and resulting spectra identified as N-cad peptides by SEQUEST based database searching. 

B Culturing dNSCs in N-cad function-blocking antibody partially rescues the ability of dNSCs to colonize the ICM. Integration frequency within the ICM was quantified (ES n= 45, dNSC n= 25, dNSC + N-cad antibody n= 83). C Sample image of dNSC aggregation (scale bar =50μm). DI-IV Representative optical sections (spaced by 10μm in a single confocal Z-stack) confirming that blocking N-cad in dNSCs rescues ICM colonization.
had selective inhibitors which were readily available. To assay whether enrichment within the neural populations predicted functionally relevant molecules, we compared the colony-forming ability of ES cells to each of the pNSC and dNSC populations. Colony formation in each inhibitor was normalized to the appropriate vehicle control at a concentration of inhibitor previously established to be selective for the target [351-355].

Of the proteins enriched on the cell surface of NSCs compared to ES cells, we chose inhibitors directed against Erbb2, Hedgehog signalling, EphA4, N-Cad and c-Kit. Aside from the N-cad function blocking antibody, each of these had significant effects on at least one of the ES-derived NSC populations without affecting ES cells ($F_{3,15} = 8.27$, $p<0.05$). Inhibition of Erbb2 blocked pNSC colony formation ($p<0.05$), but did not interfere with proliferation of ES cells or dNSCs (Fig 6A). Inhibition of hedgehog signalling blocked formation of both pNSCs and dNSCs without affects on ES cells ($p<0.05$) (Fig 6A). EphA4 inhibition blocked NSC colony formation without affecting ES colony growth ($p<0.05$) (Fig 6A). Conversely, N-cad inhibition had no effect on colony formation of either pNSCs or dNSCs ($p>0.05$), whereas c-Kit inhibition increased colony formation of pNSC but did not affect dNSCs ($p<0.05$) (Fig 6A). We then asked whether the inhibitors were blocking signalling involved in viability or proliferation. Erbb2 inhibition did not have an effect on pNSC viability but limited division; these data indicate that Erbb2 signalling is required for proliferation (Fig 6B). Conversely, EphA4 inhibition lead to increased cell death of both pNSCs and dNSCs (Fig 6B), suggesting that signalling through EphA4 is required for NSC viability. Hedgehog signalling mediates ventricular-derived
Figure 6. The ES-derived neural lineage predicts molecules that are functionally relevant for NSCs derived from adult mice. A Proteins found to be highly enriched in either pNSCs or dNSCs relative to ES cells were inhibited with Erbb2 inhibitor II, Cyclopamine (Hedgehog signalling), Gleevec (c-Kit) or an EphA4 blocking peptide. The frequency of 50µm colony formation was then compared between the different populations after normalizing to an appropriate vehicle control for each inhibitor. The normalized frequency of colony formation is depicted ± s.e.m. An ANOVA was used to assess significance ($F_{3,15} = 8.27, p<0.05$). The Erbb2 inhibitor blocked pNSC formation without impairing either ES or dNSC colony-formation, $p<0.05$. Cyclopamine blocked the formation of both pNSCs and dNSCs ($p<0.05$), as did the
EphA4 inhibitor peptide \( p < 0.05 \). Conversely, Gleevec enhanced pNSC formation \( p < 0.05 \), but did not affect colony formation of either ES or dNSCs. Notably, inhibitors which blocked the formation of ES-derived NSCs, also blocked the formation of adult mouse derived NSCs. The Erbb2 inhibitor, Cyclopaamine and the EphA4 inhibitor peptide all blocked aNSC formation \( p < 0.05 \). In the case of Gleevec, it impaired aNSC formation, in contrast to its lack of effect on ES-derived dNSCs and promotion of pNSC formation. B EphA4 and Hedgehog signalling have significant effects on NSC survival, \( p < 0.05 \). Viability after 24hrs and is depicted \( \pm \text{s.e.m.} \).

Viability in Erbb2 inhibitor and Gleevec was unaffected. C pNSC colony size (depicted \( \pm \text{s.e.m.} \)) was quantified in the presence of 4uM Gleevec and no difference was apparent relative to the vehicle control, indicating that the increase of pNSC colony formation in the presence of Gleevec is not the result of increased proliferation.

dNSC proliferation [356]. We found that hedgehog inhibition reduced pNSC viability and prevented division, suggesting that hedgehog signalling mediates both NSC viability and proliferation (Fig 6B). c-kit inhibition did not have an appreciable affect on viability or the colony size of pNSCs (Fig 6B,C), suggesting that c-kit signalling may regulate pNSC quiescence. To ensure that these results were due to specific inhibition of the proteins of interest, we used siRNA to selectively limit protein abundance of each. Examples of the efficiency by which transcript abundance is diminished are provided for EphA4 (Fig 7B) and c-Kit (Fig 7C).

Selective inhibition using siRNA produced the same effect as use of small molecule inhibitors (compare Fig 7A with 7D). These observations are summarized as a model in Fig 8.

We then asked whether the molecules that affected the ES-derived NSCs were relevant to murine forebrain-derived populations. Using the same doses that affected ES-derived NSCs but did not impair ES colony formation, 3 of the 4 inhibitors blocked colony formation from forebrain derived NSCs, supporting the relevance of predictions made with the \textit{in vitro} model to the \textit{in vivo} case (Fig 6A). Indeed, while hedgehog signalling is an established mediator of NSCs \textit{in vivo}, this work implicates Erbb2 and EphA4 in the biology of adult mouse-derived NSCs.
Figure 7. Independent confirmation that inhibition of the molecules of interest is responsible for the effects on NSC function. To confirm that the effects observed on NSC function were not due to off-target effects of the small molecules, we used siRNA to selectively knockdown abundance of the molecules of interest. A Reproduction of the data obtained using small molecules (Fig 6A) for the purposes of comparison. B,C Comparison of transcript abundance in samples treated with EphA4 (B) and cKit (C). Relative abundance was measured using quantitative PCR and referenced against samples treated with Scramble siRNA. D Colony formation of pNSCs and dNSCs in the presence of siRNAs against EphA4, Erbb2 and cKit. The frequency of 50µm colony formation was compared, normalizing each to the result using control Scramble siRNA. The normalized frequency of colony formation is depicted ± s.e.m.
Discussion

Direct glycoprotein profiling of NSCs bypassed the incongruency inherent in making inferences about protein abundance from transcriptional profiling [334], and revealed novel regulators in this lineage. The veracity of this model and profiling approach is demonstrated by markers presently used to enrich neural stem/progenitors from mice (prominin and CD24) being enriched in dNSCs relative to ES cells. This approach has contributed to the identification of several molecules for which we have experimentally demonstrated biological roles in the earliest NSC populations, which suggests that the remainder of the profile is also likely to be rich in mediators of NSC behaviour.

The data from chimeric mice indicate that pNSC progeny are biased towards the germinal zone of the brain and away from cardiomyocytes compared to ES cells. Ultimately however,
both pNSC progeny and ES progeny were distributed to tissue types throughout the chimeric embryos, indicating that pNSCs are lineage specified but uncommitted. The lineage bias of pNSC progeny in chimeric mice are similar to recent reports that iPS cells derived from different somatic tissues can retain pluripotency, but still be biased to the somatic tissue of origin in differentiation assays in vitro [357]. Clearly, the retention of pluripotency is not inconsistent with the maintenance of differentiation biases.

Our data support a model where N-cad is upregulated upon loss of Lif/Stat3 signal transduction, altering the adhesive profile of dNSCs such that they are incompatible with integration in the ICM. N-cad is expressed in several tissues within the developing embryo, including the neural plate from ~E7.5, consistent with an early role for N-cad in organization of neural cell types. It is a member of a family of calcium-dependent cell-adhesion molecules that preferentially form homotypic bonds that are important in maintaining tissue organization [358-360]. The observation that pNSCs, cultured in SFM + Lif, maintain a comparable ability to colonize the ICM as ES cells indicates that loss of BMP plays a minor role, if any, in the upregulation of N-cad. Furthermore, while dNSCs cultured in FGF-containing media almost entirely lose the ability to colonize the ICM, addition of Lif to dNSC culture conditions results in a partial rescue of ICM compatibility. These data, combined with the observation that addition of Lif to dNSC culture does not maintain E-cad levels equivalent to that of ES cells suggests that Lif signal transduction is critical to maintaining the low N-cad levels required to colonize the ICM. That the Lif based rescue of dNSC compatibility is only partial suggests that FGF and/or other signalling may also be involved in the adhesive profile transition. Notably, a combination of both Lif and BMP appears to induce ICM competence or reversion from more mature cell types [361, 362]; our data suggest that Lif-mediated activation of JAK/STAT and/or MAPK
signalling cascades is responsible for a decrease in expression of N-cad that is not compatible with ICM colonization.

Hedgehog signalling is crucial to neural development, mediating morphogenetic movements, as well as specifying differentiation of neural subtypes [363-365]. As far as NSCs are concerned, Hedgehog signalling is required for the proliferation of hippocampal adult neural precursors [363] as well as of subependymal dNSCs [364]. Cyclopamine –an inhibitor of hedgehog signalling- was tested as an inhibitor on the basis that NSC populations had >10 times the MS signal of hedgehog signalling components HhIP and Smo [366] that ES cells do. HhIP is a pan-hedgehog inhibitor expressed by E10.5 [367], while Smo is expressed by E7.0 of development and in the neural folds by E9.0 [368]. Cyclopamine is known to block the proliferation of hippocampal adult neural precursors [363] as well as of subependymal dNSCs [364]. Given this, the blockade of neurosphere formation via hedgehog inhibition effectively served as a positive control to confirm that this in vitro model can distinguish neural-specific signalling in both pNSCs and dNSCs from the background case of pluripotent ES cells.

Selective inhibition implicated several other candidates from the MS profiling in aspects of NSC function. Our data indicate that inhibition of c-Kit increases the frequency of colony-forming pNSCs but not ES-derived dNSCs or adult dNSCs. One possibility is that c-kit mediates quiescence of pNSCs, perhaps as regulator of the pool size of the pNSC population. Inhibition of EphA4 in ES-derived NSC populations indicates that it is required for the viability of the earliest NSC populations. This result was replicated in adult-derived NSCs, confirming that ephrin signalling through EphA4 is required for viability of all murine NSC populations. Indeed, these observations are in agreement with the independent conclusion that Ephrin B3 signalling through Epha4 promotes cell viability in the adult subependyma, whereas its absence leads to an
increased apoptosis [369, 370]. Erbb2 signalling blocks colony formation of ES-derived pNSCs as well as adult dNSCs. Erbb2 ligand Nrg-1 is required for proliferation of neural precursors [371], although the specific requirement for Erbb2, as opposed to another Nrg-1 receptor Erbb4, was not shown previously.

This study demonstrates the value of using CSC technology on viable cells and sequential, clonal precursor populations as a background to ascertain biologically relevant changes in cell-surface protein profiles during lineage maturation. We demonstrated functional roles for four proteins showing increased abundance during the maturation of ES cells into NSCs. These represent a small fraction of the 112 peptides in this dataset that showed significant (p<0.05) enrichment (5 times). Such confirmations, along with detection of high enrichment in known neural precursor markers such as prominin (95 times dNSC/ES) and CD24 (25 times dNSC/ES), suggest that the remainder of the dataset also will be rich in proteins relevant to NSC biology, and the earliest events in neural induction. Fruitful approaches to screening this list of candidates might include antibody-based selection to test for enrichments in colony-forming activity within the brain germinal zone populations or the use of corresponding siRNAs to determine molecules required for viability and proliferation. As such, this dataset is a step towards a higher resolution understanding of NSC biology.
Chapter 4: Global survey of genomic imprinting by transcriptome sequencing.

I conceived of the project, collected the data, co-authored the manuscripts, and contributed to experimental design and analytical approaches. All of the large scale genomic analysis was performed by Tomas Babak.

This chapter has been published as follows:

Abstract

Comprehensive monitoring of genomic imprinting requires genome-wide detection of allelic expression. We used selective priming and parallel sequencing to measure allelic bias in whole-transcriptomes of E9.5 embryos. By distinguishing sex-specific from strain-specific bias in reciprocally crossed mouse embryos, we constructed a genome-wide map of imprinted transcription. This map was able to objectively locate over 80 percent of known imprinted loci. Although more than half of all imprinted single nucleotide polymorphisms (SNPs) did not overlap previously discovered imprinted transcripts, many were associated with known clusters and likely extensions of known imprinted transcripts or ncRNA within known loci. We estimate that six novel imprinted genes were detected and confirmed with this approach. This discovery rate is consistent with existing estimates of approximately 200 murine imprinted genes. In contrast, alternate analysis of comparable transcriptome data uncovered parent-of-origin allelic effects at more than 1300 loci in the developing brain and two adult brain regions, including hundreds present in only males or females. Our independent replication of the embryonic brain stage, where the majority of novel imprinted genes were discovered and the majority of previously known imprinted genes confirmed, resulted in only 12.9% concordance among the novel imprinted loci. Further analysis and pyro-sequencing-based validation revealed that the vast majority of the novel reported imprinted loci are false-positives explained by technical and biological variation of the experimental approach. We show that allele-specific expression (ASE) measured with RNA-Seq is not accurately modeled with statistical methods that assume random independent sampling and that systematic error must be accounted for to enable accurate identification of imprinted expression. The results emphasize the importance of independent validation, and suggest that the number of imprinted genes is much closer to the initial estimates.
Introduction

Why diploid organisms would forgo the safety-net of a redundant genome and preferentially express one allele in a parent-of-origin dependent manner has been a matter of debate since the discovery of imprinted transcription. Genomic imprinting is common to eutherian mammals, although the location and makeup of imprinted regions is not always identical between species. Lack of an effective high-throughput screening approach hinders characterization of genomic imprinting. Our understanding of this issue as well as the range of processes affected by imprinting is dependent on our catalog of the identity, function, and spatial-temporal specificity of imprinted genes.

Imprinting was initially characterized with genetics. Regions where uniparental disomy is not tolerated were mapped by intercrossing reciprocal translocations. In this scheme, progeny from a single parent that has contributed both the original and translocated segment are assessed for viability (or another feature) and compared to progeny which inherited each copy of the segment conventionally, one copy from each parent [372]. Continued use of complementation tests in function-based screens provided the most dramatic examples of parent-of-origin effects on development, implicating imprinted transcription in neonatal behaviour (such as suckling [373]) as well as organism growth both pre- and post-natally [374]. Imprinting was then formally demonstrated using nuclear transfer experiments, as post-implantation development of embryos with two female pronuclei failed [375]. Poor development of extra-embryonic membranes and trophoblast suggested imprinting of genes required for development of these lineages [375]. Subsequently, the first individual imprinted transcripts were mapped [376-379]. After screening many translocations, genome-wide estimates stood at 100-200 imprinted transcripts [380].
Although based on what we now know was an overestimate for the total number of genes (60,000-100,000) and an underestimate of the number of known imprinted clusters [380, 381], the ~20 year-old estimate has endured all screening methods applied, including those that do not depend on overt phenotypes. Some of these screens revealed imprinting of transcripts that do not influence organism growth [248, 251, 382, 383]. By Dec 2008, genetic and molecular screening efforts combined yielded 128 confirmed imprinted genes in the mouse. No approach yet reported has demonstrated sufficient sensitivity and specificity to comprehensively map imprinting in a practical manner. Such an approach will accelerate our understanding of gene expression by allowing researchers to study imprinting without restricting themselves to individual loci or specific transcripts.

Transcriptome sequencing of F1 mouse hybrids provides an unbiased alternative for discovering imprinted transcription in wild-type animals [384, 385]. The approach is based on detecting allelic expression with RNA sequencing reads that map over heterozygous SNPs, where the identity of the base is used to distinguish allelic origin and a reciprocal cross is used to discriminate parent-of-origin from strain-specific biases. By screening for replicated allele bias toward the maternal or paternal allele in both crosses, we directly assessed imprinting in a wild-type setting. Our initial analysis was of sequenced whole transcriptomes from eight reciprocally crossed CAST/EiJ x C57BL6J embryos, four independently sequenced embryos from each cross. This approach yielded six novel imprinted transcripts.

However, two recent studies used this approach to identify more than 1,300 imprinted loci, including 484 ncRNAs and 347 genes that were sex-specific [386, 387]. These 1,300 loci are an aggregate of the discoveries from E15 brains, adult medial prefrontal cortex (PFC) and adult preoptic area (POA), and represent a ten-fold increase over previously known imprinted
genes. The authors suggest that improved sensitivity from increased sequencing depth and improved resolution from sequencing the parents for de novo identification of SNPs enabled these advances.

To investigate the biological robustness of these novel imprinted loci, we repeated the embryonic brain screen. Despite faithful technical reproduction of the experimental design, library construction, sequencing, and analysis, we could not reproduce the majority of novel imprinted genes. In this study we demonstrate that biological variation in the approach and technical variation of the assay introduce considerably more noise than was appreciated previously. We develop methods to account for this variation and demonstrate their utility through reanalysis of the published data mentioned above as well as new embryonic brain data.

**Materials and Methods**

**Mouse crosses**

All mice were obtained from Jackson laboratories (Bar Harbor, ME) with the exception of E9.5 Cast/Ei embryos which were obtained from The Mary Lyon Centre, (Harwell, UK), and housed in accordance with the Institutional and Governmental Animal Care Committee guidelines of the University of Toronto. For staging, embryos were assumed to be 0.5 dpc on the morning a vaginal plug was found. At E9.5 or E17.5, mothers were sacrificed by cervical dislocation and the embryos placed in RNALater (Ambion AM7020) immediately following dissection and snap-frozen in liquid nitrogen within 30 minutes. Brains were placed in Trizol and immediately snap-chilled in liquid nitrogen. Four E9.5 embryos from each cross were collected along with two male BxC E17.5 brains, 1 female BxC brain and 1 male CxB brain.
Genomic library construction and sequencing

Whole embryos and brains were homogenized in Trizol (Invitrogen), RNA was extracted according to the manufacturer's instructions, and integrity was confirmed on an Agilent Bioanalyzer. A detailed description of whole-transcriptome amplification used for preparation of the libraries from E9.5 embryos is described in Armour et al (2008). In brief, 1 ug total RNA from each embryo was reverse-transcribed with Superscript III (Invitrogen) in a 20 uL reaction containing 2 mM dNTP, 5 mM DTT, 1 unit RNase OUT, and 10 uM primers (see Armour et al (2008) for sequences), incubated 40°C for 90 minutes, 70°C for 15 minutes, and cooled to 4°C. RNA was degraded by addition of 1 uL RNase H (Invitrogen) and incubated at 37°C for 20 minutes then heat-denatured at 75°C for 15 minutes. Following PCR purification (Qiagen), first-strand product was subjected to second strand synthesis in a 100 uL reaction containing 10 uM antisense primers, 16.5 units Klenow (NEB: M0212L), and 0.2 mM dNTP by incubating 37°C for 30 minutes. Following PCR-purification (Qiagen), adaptors and priming sites were added through PCR using Roche High-Fidelity system at recommended conditions for 35 cycles (94°C 15", 60°C 30", 72°C 1'+10''/cycle) (Illumina). To ensure initial primer annealing, annealing temperature for the first two cycles was reduced to 40°C for 2 minutes. Following PCR purification (Qiagen), library integrity was verified by resolution on 2% agarose. Eight whole-embryo libraries were sent to Illumina (Hayward, California) for sequencing at 256 Mb/embryo. Of 78,408,722 reads (average 9.8 million reads/embryo; range: 8.2 million to 11.7 million), 58,693,820 (74.8%) aligned as described below.

Libraries from the E17.5 brains were prepared as follows. PolyA+ RNA was isolated using a
Dynabead mRNA Purification Kit according to the manufacturer's instructions (Invitrogen).
Double stranded cDNA libraries were made using the Illumina mRNA-Seq kit according to the manufacturer's recommendations (Illumina) and converted into libraries (adapter ligation, PCR, cleanup) using a NEBNext kit according to the manufacturer's recommendations (NEB). We retained strand-specificity by using dUTP during second-strand synthesis and an UNG treatment prior to the final amplification as described previously [388]. Libraries were verified on an Agilent Bioanalyzer and by qRT-PCR, and each sequenced to 36 bp on an Illumina HiSeq 2000 platform, yielding on average 80.1 million reads. Raw data and alignments are available at GEO under accession GSE27016. Brains were processed independently such that each library is derived from a unique sample without pooling.

**Genome alignment of E9.5 embryo transcriptomes**

All Illumina reads were aligned with Eland (Illumina) against mouse NCBI genome build 36 (mm8) and against UCSC known genes [389] and mouse ESTs (mm8 release) to capture alignment to splice-sites. Reads with unique genome coordinates (i.e. mapped to only one location) were retained in addition to redundantly mapped reads that overlapped unique reads at only one genomic location (presumably emerging from the same transcript).

**SNP identification within E9.5 transcriptome data**

Eland aligns reads with up to two genomic mismatches and provides quality scores for each base. In order to calibrate our SNP-calling threshold we analyzed SNP positions previously identified by the Perlegen-led study [390]. We added Phred-like scores for identical bases and assessed true and false calls by agreement with this previous study. Scores lower than 20 were treated as 0.
To rule out additional errors arising from PCR, mispriming, and sequencing, we also required that mismatches be supported by at least two biological replicates (embryos or adult brain were treated independently), did not consider mismatches within the first seven bases of each read, and required that the alternate base exceed 10% in frequency. We used a combined-Phred of 80, corresponding to a FDR of 3.8% to detect 160,078 SNPs in the transcriptome. 3.8% is likely an overestimated FP rate given that the Perlegen study also had a FP rate of ~3% [391]. The false-negative rate is more difficult to ascertain because of the large variance in sequencing coverage. The achieved level of 64% recall for Perlegen-identified SNPs suggests that we are missing 36% of SNPs, although the majority do not make the threshold due to low sequence coverage.

Quantifying parent-specific expression within E9.5 transcriptomes

Illumina reads were split by cross and aligned as described above. Ratios of CAST/EiJ to C57Bl/6J allele counts were determined for all identified SNPs. The cumulative binomial distribution was used to assess the probability that observed ratios occurred by chance. To correct for minor alignment bias against CAST sequences, we set the probability of success to reflect the overall ratio of C57Bl/6J to CAST/EiJ (0.42). For all analyses an imprinting score was determined as the log10 of the less significant binomial p-value of the two crosses, where both crosses agree on direction (either paternal or maternal allele counts being higher in both crosses). Allele counts were tallied independently, or grouped by transcript coordinates or genomic windows. A genomic window of 10kb was used because ~30 allele calls are minimally required to reliably assess allelic bias, and the mean allele count in a cross per 10kb was 37.

Transcript definition
Unless otherwise indicated, "transcripts" in analysis of E9.5 embryos comprise mouse UCSC known genes [389] and mouse GenBank mRNAs that do not overlap UCSC known genes on the same strand (mm8; UCSC download). Transcripts mapping to the sex chromosomes and mitochondrial chromosome were not considered.

**Calling imprinted genes from brain transcriptomes**

We downloaded sequencing data and SNP calls from GEO (GSE22131) and used the same strategy as Gregg et al. [386, 387] to score imprinted expression. We aligned the data to UCSC known genes (transcripts) and ncRNAs downloaded from the Functional RNA Database [392]. We called a SNP 'imprinted' if greater than 50% of the reads in each sample mapped to alleles from the same parental sex ($p<0.05$, chi-square test) in both samples. We compiled a non-redundant set of gene models from UCSC known genes [393] by taking the union of transcribed bases when multiple isoforms overlap (i.e. collapse isoforms into one gene model that contains all exons), yielding 26,214 models, of which 19,867 were coding (available under GSE27016). 13,604 gene models had at least one SNP identified by Gregg et al [387]. We called a gene imprinted if it contained one or more imprinted SNP(s). Sex-specific calls were made as described [386].

**Mock comparisons**

Imprinting calls were made using criteria described above. To minimize sampling bias reads aligned to chromosome X and the mitochondrial chromosome were not considered in this analysis. Furthermore, all samples were normalized to have the same number of aligned input reads. For example, for the PFC comparison, the BxC male sample had the fewest aligned reads.
so reads were randomly removed from the three remaining samples (BxC female, CxB male, and CxB female) to exactly match BxC male.

Discovery of novel imprinted genes in brain transcriptomes
For global discovery of novel imprinted genes we aligned to gene models as described above, all possible splice-junction sequences (100 bases; 50 bases from each flanking exon) representing up to two exon-skipping events in these gene models, as well as the genome. We converted all alignments to genomic coordinates and retained uniquely mapping reads for further analysis. 79.9% of reads aligned uniquely (on average 80.1% per sample), of which 7.7% spanned splice-junctions (17.6% of reads that overlapped coding exons also spanned splice junctions) and 0.12% represented exon-skipping events. Imprinting was assessed at each SNP as described above.

Confirmation of E9.5 imprints
In a 20 ul reaction, 80 ng of pooled embryonic RNA (pooled across 4 littermates) for each of the reciprocal crosses was amplified using the OneStep RT-PCR kit according to the manufacturer’s instructions (Qiagen, 210212). All primers were added to a final concentration of 0.4 nM and cycling followed: 30 min at 50°C, 15 min at 95°C, [30 sec at 94°C, 30 sec at 54°C, 1 min at 72°C] x 35 cycles and 10 min at 72°C. Half of the reaction was digested with the enzyme listed in Table 1 according to the manufacturer’s instructions. Half remained undigested. Reactions were visualized on 1.0% agarose gels using Sybr-Safe (Invitrogen, 33102). A complete list of primers, amplified regions including target SNPs, and restriction enzymes are available in Table 1.
Table 1. Primer pairs used in cDNA-RFLP and RT-PCR Sanger confirmations.

<table>
<thead>
<tr>
<th>ID</th>
<th>Forward Primer</th>
<th>Reverse Primer</th>
<th>Enzyme (if RFLP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zabc2fy714830 (RFLP)</td>
<td>ATGACACAGCTTTTTTGGAGCGGCGCTG</td>
<td>CAGCGCTGAGCTTTTCTTTTG</td>
<td>HpyCHIV</td>
</tr>
<tr>
<td>Zabc2fy714830 (Sanger)</td>
<td>GAGCCGCTTTTTTTTCTTTCTTTG</td>
<td>GCAAAAGGTTTCTGTCAAAGGA</td>
<td>HindIII</td>
</tr>
<tr>
<td>D2ferd173e</td>
<td>TACGCTGCTCCCTTTGAGTA</td>
<td>ATACCGCTGGGACGACGAGA</td>
<td></td>
</tr>
<tr>
<td>nkGA40265</td>
<td>AAAGCCACACACTTGTTCA</td>
<td>TACGGGAGTCACTACACTTCC</td>
<td></td>
</tr>
<tr>
<td>Cndl</td>
<td>CTAACGCCCAGCTTTTGTGCG</td>
<td>GCTCTGATGTCTGTCTG</td>
<td></td>
</tr>
<tr>
<td>A25007F06Rk</td>
<td>ATCAGCCACGATTTCCAAC</td>
<td>TGCTGGCTTTGAAGGTGTA</td>
<td>Haelll</td>
</tr>
<tr>
<td>A35007H08Rk</td>
<td>ACTTCCGCTGACTGTTTCAAC</td>
<td>GCAAGGCTTTTACTACCTTCA</td>
<td></td>
</tr>
<tr>
<td>EG330552</td>
<td>CATGGAAAAATCTGACACACCTACCTCC</td>
<td>GCTGTTGATAGGAGGCGCTCAA</td>
<td></td>
</tr>
<tr>
<td>Frel5-enc</td>
<td>ATTTCCACCTTGTGCTTCCCT</td>
<td>CACCTTTGCGAGGGCAAAGGG</td>
<td></td>
</tr>
<tr>
<td>NMIT3 (RFLP)</td>
<td>ATACACGCTGCTGGAGGAGGCTC</td>
<td>AGATCGCTGGTCTGAGAT</td>
<td>Acil</td>
</tr>
<tr>
<td>NMIT3 (Sanger)</td>
<td>ATGCTAGGTCCTTTGGAGGCG</td>
<td>CTGCAACACCCAACCTGAG</td>
<td></td>
</tr>
<tr>
<td>AK018142</td>
<td>GGTGTAAGATGCTGAGGTTAGGGTTG</td>
<td>TGCGCAAGAGGTGATCTGATTAC</td>
<td></td>
</tr>
<tr>
<td>AK046138</td>
<td>GTACGACGAAAGGTTGACAGGT</td>
<td>CAAGGTGCGGACCTAATAAGA</td>
<td></td>
</tr>
<tr>
<td>AK148555</td>
<td>GACAGGCTGCTGATTGCTGCTA</td>
<td>CTTTGCTTTTTTCAAAGA</td>
<td></td>
</tr>
<tr>
<td>AK086774</td>
<td>AGAATGCTAAAACTCCCTGCA</td>
<td>GAGATCTTTTCTTTCTTTG</td>
<td></td>
</tr>
<tr>
<td>AK030713</td>
<td>CACAGTGGAGGAGGAGGTAG</td>
<td>GAAATCTCTCCTGGACACCA</td>
<td></td>
</tr>
<tr>
<td>AK141557</td>
<td>ATGCAACACCTGAAAGGCA</td>
<td>GATACACATCTGCTATGGA</td>
<td></td>
</tr>
<tr>
<td>NMIT2</td>
<td>CAATATGAAGAGCAAGAGCGGG</td>
<td>AGACAGGTCTCTGCGAGGAGA</td>
<td>BstXI</td>
</tr>
<tr>
<td>NMIT1</td>
<td>TGTTGCTAGGACATGACATAT</td>
<td>GAGGAAATACGGGCGCAAGGA</td>
<td>Acil</td>
</tr>
<tr>
<td>Nilet4</td>
<td>GGTCTACCTGACCTGCGG</td>
<td>GCTGAAATGAGCAAGAAAAGCA</td>
<td></td>
</tr>
<tr>
<td>NMIT4</td>
<td>TGGCTAGGATATAGGAGGACGG</td>
<td>TGTTACGAGGGTCTGAGGAGG</td>
<td></td>
</tr>
<tr>
<td>72s112</td>
<td>TGCCACAGTTGCTTGTAAATCA</td>
<td>AGATCGACAGGCTTCTCTTC</td>
<td></td>
</tr>
</tbody>
</table>

Pyro-seqencing validation of E17.5 imprints

We identified SNPs amenable to pyro-seqencing (non-consecutive nucleotides separated from other SNPs) within transcripts of interest and designed assays using Pyromark Assay Design 2.0 requiring assay scores >87 (Table 2). The sequence ‘CGCCAGGGTTTTTCCACGAC’ was added to the 5’ end of all primers designated for biotinylation to enable biotin incorporation during PCR following an approach reported previously [394] with some modifications. Specifically, biotinylated amplicons were generated directly from RNA using the Pyromark OneStep RT-PCR kit (Qiagen) according to the manufacturer’s instructions with the addition of a third HPLC purified universal biotinylated primer (biotin-CGCCAGGGTTTTTCCACGAC) added at 9/10 the recommended molarity to supplement the RT-PCR primer designated for biotinylation, which was added at 1/10 the recommended molarity. Other combinations ranging from 5:5 to 9:10 did not result in a
noticeable difference in product yield and size (as assayed by an Agilent Bioanalyzer) or pyrosequencing performance. We performed sequencing on a Pyromark Q96 MD (Qiagen) according to the manufacturer’s instructions and quantified allelic bias using the AE quantification software included with the instrument. To maximize sensitivity of detecting small allelic changes we defined the null ratio by sequencing DNA as described previously [395]. Transcripts were called imprinted if the difference in allelic bias between the reciprocal samples were >5.02% and where each ratio was reciprocally biased (i.e. in opposite directions) relative to ratio obtained with DNA. 5.02% represents 2 standard deviations of variance determined from comparing allelic ratios between biological replicates (Figure 1) and corresponds to a significance of $p<0.022$.

Both biological replicates had to meet these criteria and DNA ratios were averaged between the two DNA samples prior to normalization. A negative pyro-sequencing call includes cases that do not meet significance as well as cases where reciprocal bias was not observed. The majority of negatives fall into the latter category and relaxing the threshold of significance to $p<0.2$ did not change the outcome of any calls. 5/42 assays attempted were removed due to technical failure (two because of a high skew in the DNA ratio and three where the Pyromark AE quantification software could not accurately identify peaks) leaving 37 assays successfully executed across all 5 samples.

<table>
<thead>
<tr>
<th>Gene/ Transcript</th>
<th>F1 primer</th>
<th>R1 primer</th>
<th>Seq primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peg3</td>
<td>CCGCAGGGTTTTCCCACTACGACATG GCGAAGATCATTATGTAAGG</td>
<td>AGTGAACAAATTCCAAGACAGAG</td>
<td>AGAGTTGCCCAAGT</td>
</tr>
<tr>
<td>Rasgrf1</td>
<td>CCGCAGGGTTTTCCCACTACGACAAAAAGAAGTCCAGAGAGG</td>
<td>TGGTGGGGAATTTAATCTGGTG</td>
<td>GTCCTCATCACTGGTT</td>
</tr>
<tr>
<td>Frat3</td>
<td>CCGCAGGGTTTTCCCACTACGACATATCACAGAACTGCAGCTGTTCC</td>
<td>TGGGAAGCTTCTCTATATGGG</td>
<td>GGACACAGTCACACCAAC</td>
</tr>
<tr>
<td>Gnas</td>
<td>CCGCAGGGTTTTCCCACTACGACCAAGCCAAGGAGG</td>
<td>TGGCCTCTGGTCTTGCA</td>
<td>GCGGGGCTGCTGTTT</td>
</tr>
<tr>
<td>DOKist4</td>
<td>CCGCAGGGTTTTCCCACTACGACAGGG</td>
<td>CTCCTCTGGCTCCCCCTCAT</td>
<td>TGGCTCCCCCTCTAT</td>
</tr>
<tr>
<td>U80983</td>
<td>CCGCAGGGTTTTCCCACTACGACACAGAGCAAGAAGCAAAATAGAAGTC</td>
<td>TTCCTCTCAGCATCTAAGAG</td>
<td>CATACGAGCTGATCACCC</td>
</tr>
<tr>
<td>Cdc40</td>
<td>CCGCAGGGTTTTCCCACTACGACACAGGG</td>
<td>TTCCTCTCAGCATCTAAGAG</td>
<td>CTCCTCTCAGCATCTAAGAG</td>
</tr>
<tr>
<td>Bot211</td>
<td>CCGCAGGGTTTTCCCACTACGACAGGG</td>
<td>TGGTCTCTGCCCCACCTA</td>
<td>TGGCTAGGAGCCCCTT</td>
</tr>
<tr>
<td>Gene</td>
<td>Primers used to validate or refute putative imprinted loci by pyro-sequencing. Table provides the Gene/Transcript, as well as F1, R1, and S1 primer sequence.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mapt</td>
<td>CGCCAGGGTTTTCTCCAGTCAGCGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adam23</td>
<td>CGCCAGGGTTTTCTCCAGTCAGCGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wars</td>
<td>CGCCAGGGTTTTCTCCAGTCAGCGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pld4</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cpx2</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhodl3</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nsg1</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malat1</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Herc3</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apba2</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limd1</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syt16</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trak2</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bmp1</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S230400G2</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hsp90ab1</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nme1</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cald1</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nsg2</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK138412</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ddi4</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dvl3</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ina</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1700027N1</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0Rik</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pitpm2</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK011885</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asb6</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chmp1a</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enpp5</td>
<td>GTGGACAGAGGAGAGAGA CGCCAGGGTTTTCCCAGTCACGACGACGAG ATGGATGGCTGAGT GAGTGGCTGAGT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

SNP calling within the E9.5 transcriptome

We chose to sequence day 9.5 embryos from a reciprocal cross of the CAST/EiJ (C) and C57BL/6J (B) strains of mice. E9.5 is a widely interrogated stage of development and consequently an ideal benchmark for imprinting studies [251]. We selected the divergent castaneus strain to maximize single nucleotide polymorphisms that could be used for inference of allelic expression. Technical issues with the analysis of total RNA have restricted most researchers to the study of polyadenylated transcripts, consequently excluding a considerable fraction of ncRNAs that are coincidentally the fastest growing class of imprinted transcripts. So we generated strand-specific cDNA from total RNA primed using hexamers devoid of ribosomal complementarity to amplify all cellular RNA except for ribosomal RNA. We then conducted
high throughput sequencing of the cDNA to obtain over 78 million reads, which we used to uncover SNP polymorphisms between strains that could be used as markers for allelic expression. SNPs were identified by combining data from both crosses and identifying mismatches to the C57Bl genome (NCBI build 36). To distinguish de novo SNPs from sequencing errors we used Solexa quality scores, calibrating our approach to the set of *castaneus* SNPs previously identified by Perlegen [390] (Figure 2, see Materials and Methods). For a threshold corresponding to a FDR of 3.8% and sensitivity of 64% (Figure 2A), we detected 160,078 expressed SNPs in the genome, 75% of which had not been previously discovered (Figure 2B). We selected this stringent threshold for SNP calling to reduce the FDR in subsequent allelic bias calls.

Figure 2. Calibrating SNP discovery. (A) Receiver operating characteristic curve based on calibration of combined Phred scores with the Perlegen data. (B) 160,078 total SNPs were identified, 40,500 that had been previously discovered.
The observed SNPs were then used to detect allelic bias characteristic of imprinting (Figure 3), with the binomial distribution used to calculate the probability of observed allelic counts varying from expected proportions. If the reciprocal crosses agreed with each other on parent-of-origin bias, we computed a conservative “imprinting score” (IS) by using the least significant binomial p value (see Materials and Methods).

Figure 3. Measuring allelic expression by detecting and counting SNPs. Schematic of approach used to detect and count SNPs. Reads were aligned to the genome. SNPs were distinguished from sequencing errors by setting Phred score thresholds, calibrating to a set of previously discovered SNPs [390]. In addition to treating SNPs individually, SNPs could also be pooled by 10kb windows or transcript coordinates to improve statistical discrimination. An imprinting score was calculated from the reciprocal cross data by using the binomial distribution.

### Calling E9.5 Imprinted Genes with a Sliding Window

To generate a genome-wide map of imprinting, we scanned the genome for clusters of imprinted SNPs, using a 10 kb sliding window in which allele counts were grouped together to take advantage of proximally imprinted SNPs. The sliding window method readily detects maternal-specific expression of the mitochondrion and paternal-specific expression of the Y-chromosome, as expected (Figure 4). Furthermore, we
Figure 4. Genome-wide identification of imprinted regions. Imprinting scores were computed for allele counts grouped in 10 kb windows every 2 kb across the genome. Scores were arbitrarily set to negative for paternal bias. Regions are referenced by first reported imprinted transcript, or by common literature use. Imprinting on the X chromosome was assessed using data from two reciprocally crossed female embryos to avoid male maternal bias.

correctly located 14 of 17 known imprinted loci [381, 396, 397] with $|IS| > 3$ (p<0.001). On the other hand, new loci were not apparent by this method, suggesting that most of the major imprinted loci that are expressed throughout the E9.5 stage are already accounted for.

Calling E9.5 Imprinted Genes using Transcript Boundaries

To further improve sensitivity and to capitalize on the assay's ability to detect strand specificity, we used transcript annotations to combine SNPs and assess the likelihood of allelic expression.
As before, combining of allelic counts reduces variance, allowing a more accurate assessment of allelic expression. The p-values of transcripts for each cross are plotted in Figure 5A,B. Points along the positively sloping diagonal correspond to transcripts with strain-biased expression (upper right being CAST/EiJ-specific), and comprise the overwhelming majority of allele-biased expression, underscoring the importance of carrying out the reciprocal cross to distinguish imprinted effects. Parent-of-origin, sex-biased transcripts fall on the negatively-sloping diagonal and, as expected, are prominently enriched for known imprinted transcripts. Since confidence scores reflect the likelihood that a particular transcript is biased and not necessarily the magnitude of that bias, it is helpful to show the magnitudes of imprinting we observed: Figure 5C shows measurements of parental sex-specific bias made by summing parental alleles for 27 previously known imprinted transcripts as well as 12 transcripts that have not previously known to be imprinted.
Figure 5. Allele-biased expression of known transcripts. (A,B) Transcripts comprise UCSC known genes and GenBank mRNAs (outside UCSC known genes) that are expressed, do not map to mitochondrial or sex chromosomes, and contain SNPs (n=22,932). An allele-bias score
was computed from each cross independently using allele counts across all SNPs within the transcript boundary (including introns). Transcripts with biased expression toward sex-of-origin map to the negatively sloping diagonal and are enriched for known and novel experimentally confirmed imprinted genes. Transcripts biased for strain-of-origin map to the positively sloping diagonal and are enriched for independently identified cis-regulated variants (see text). (C) Parent-of-origin allele bias computed using allele counts from both crosses for selected known and novel imprinted transcripts. Transcripts are UCSC known genes [389] or MGD transcripts [398], as defined in Table 3.

<table>
<thead>
<tr>
<th>ID</th>
<th>Locus</th>
<th>Novelty Category</th>
<th>Putative Extension of</th>
<th>Expr. Origin</th>
<th>Additional Validation</th>
<th>Imprinting Score</th>
<th>Allele-Specific Reads</th>
<th>Parental Bias (Sanger)</th>
<th>Parental Bias (Sanger)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zdbf2</td>
<td>NA</td>
<td>New Gene</td>
<td>NA</td>
<td>P</td>
<td>RFLP, Sanger</td>
<td>-2.1</td>
<td>147</td>
<td>0.67</td>
<td>1</td>
</tr>
<tr>
<td>D2Ertd173e</td>
<td>Gnas</td>
<td>Extension</td>
<td>Nespas</td>
<td>P</td>
<td>Sanger</td>
<td>-4</td>
<td>61</td>
<td>0.87</td>
<td>0.95</td>
</tr>
<tr>
<td>mKIAA0265</td>
<td>Cogp2</td>
<td>Known</td>
<td>NA</td>
<td>M</td>
<td>Sanger</td>
<td>7.11</td>
<td>2902</td>
<td>0.58</td>
<td>0.67</td>
</tr>
<tr>
<td>Casd1</td>
<td>Peg10</td>
<td>Known</td>
<td>NA</td>
<td>M</td>
<td>Sanger, RFLP</td>
<td>1.78</td>
<td>743</td>
<td>0.57</td>
<td>0.62</td>
</tr>
<tr>
<td>A230057D06Rik</td>
<td>Ube3a</td>
<td>Extension</td>
<td>Zfp127</td>
<td>P</td>
<td>Sanger</td>
<td>-4.8</td>
<td>55</td>
<td>0.91</td>
<td>0.97</td>
</tr>
<tr>
<td>A330076H08Rik</td>
<td>Ube3a</td>
<td>Extension</td>
<td>Zfp127</td>
<td>P</td>
<td>Sanger</td>
<td>-2.9</td>
<td>40</td>
<td>0.93</td>
<td>1</td>
</tr>
<tr>
<td>E6330552</td>
<td>Ube3a</td>
<td>Extension</td>
<td>Zfp127</td>
<td>P</td>
<td>Sanger</td>
<td>-16</td>
<td>353</td>
<td>0.87</td>
<td>0.86</td>
</tr>
<tr>
<td>Frat3-ext</td>
<td>Ube3a</td>
<td>Extension</td>
<td>Frat3</td>
<td>P</td>
<td>Sanger, RFLP</td>
<td>-3.15</td>
<td>68</td>
<td>0.96</td>
<td>1</td>
</tr>
<tr>
<td>NMIT3</td>
<td>Grb10</td>
<td>Known</td>
<td>NA</td>
<td>M</td>
<td>Sanger, RFLP</td>
<td>17.1</td>
<td>213</td>
<td>0.93</td>
<td>1</td>
</tr>
<tr>
<td>AK018142</td>
<td>Dlk1</td>
<td>Extension</td>
<td>Meg3</td>
<td>M</td>
<td>Sanger</td>
<td>39.9</td>
<td>471</td>
<td>0.94</td>
<td>1</td>
</tr>
<tr>
<td>AK046138</td>
<td>Dlk1</td>
<td>Extension</td>
<td>Meg3</td>
<td>M</td>
<td>Sanger</td>
<td>13.37</td>
<td>110</td>
<td>0.97</td>
<td>1</td>
</tr>
<tr>
<td>AK148955</td>
<td>Dlk1</td>
<td>Extension</td>
<td>Meg3</td>
<td>M</td>
<td>Sanger</td>
<td>19.25</td>
<td>262</td>
<td>0.94</td>
<td>1</td>
</tr>
<tr>
<td>AK086774</td>
<td>Dlk1</td>
<td>Extension</td>
<td>Rian</td>
<td>M</td>
<td>Sanger</td>
<td>5.86</td>
<td>47</td>
<td>0.98</td>
<td>1</td>
</tr>
<tr>
<td>AK050713</td>
<td>Dlk1</td>
<td>Extension</td>
<td>Rian</td>
<td>M</td>
<td>Sanger</td>
<td>2.23</td>
<td>68</td>
<td>0.88</td>
<td>1</td>
</tr>
<tr>
<td>AK141557</td>
<td>Dlk1</td>
<td>Extension</td>
<td>Rian</td>
<td>M</td>
<td>Sanger, RFLP</td>
<td>30.1</td>
<td>292</td>
<td>0.95</td>
<td>1</td>
</tr>
<tr>
<td>NMIT2</td>
<td>Dlk1</td>
<td>Extension</td>
<td>Rian</td>
<td>M</td>
<td>Sanger, RFLP</td>
<td>1.04</td>
<td>17</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NMIT1</td>
<td>Dlk1</td>
<td>Extension</td>
<td>Rian</td>
<td>M</td>
<td>Sanger, RFLP</td>
<td>7.01</td>
<td>70</td>
<td>0.96</td>
<td>1</td>
</tr>
<tr>
<td>Pde4d</td>
<td>NA</td>
<td>New Gene</td>
<td>NA</td>
<td>P</td>
<td>Sanger</td>
<td>-1.9</td>
<td>108</td>
<td>0.67</td>
<td>0.64</td>
</tr>
<tr>
<td>Air</td>
<td>Igf2r-</td>
<td>Isoform</td>
<td>Air</td>
<td>P</td>
<td>qRT-PCR</td>
<td>-110.51</td>
<td>4247</td>
<td>0.76</td>
<td>NA</td>
</tr>
<tr>
<td>Tbc1d12</td>
<td>NA</td>
<td>New Gene</td>
<td>NA</td>
<td>P</td>
<td>Sanger</td>
<td>-1.8</td>
<td>287</td>
<td>0.6</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Table 3. Summary of validated novel features. Transcripts were identified from various sources as indicated, and imprinting scores were computed with the binomial distribution (see text). Additional validation comprised Sanger Sequencing of RT-PCR products and in certain cases RFLP of RT-PCR products containing SNPs within the novel imprinted feature.
Validation of E9.5 Imprinted Genes

We also used additional experimental methods to verify allelic expression of 20 individual SNPs selected to represent new aspects of the imprinted transcriptome, mainly: 1) extensions of previously known imprinted transcripts, 2) novel imprinted transcripts associated with previously known imprinted loci, and 3) previously unknown imprinted genes (Table 3). We Sanger-resequenced these regions, and, in cases where common restriction enzyme sites were available, applied classical cDNA-RFLP [378] for further validation.

The results of these validation studies show that characterization of known imprinted loci is often incomplete. A striking example of this occurs in the Grb10 locus, where extensive maternally biased transcription occurs immediately upstream and antisense to Grb10, suggesting that an imprinted noncoding transcript originates from the same promoter as the maternally expressed Grb10, but on the opposite strand (Figure 6). A noncoding transcript has not been associated with the Grb10 locus previously.

In the Dlk1 locus we observed 81 maternally imprinted SNPs within a 215 kb region that appears likely to be continuously transcribed from one strand. However, only 40 (49%) of these SNPs overlap any previous transcriptional evidence (mRNAs, ESTs, predicted coding genes), demonstrating extensive novel transcription in this locus. This transcription is consistent with extension of a Rian/Meg3 precursor that is processed into the imprinted miRNAs and snoRNAs present in this region and suggested to be coordinately regulated as a polycistronic locus [399]. We observe a continuous distribution of reads throughout the locus that is also consistent with the histone methylation patterns observed by chromatin immunoprecipitation in ES cells [400].
Figure 6. Imprinting beyond documented transcriptional boundaries. (A) UCSC genome browser view of novel imprinted transcript (NMIT3) with no previous transcriptional or imprinting evidence. X-axis defines position in genome. Transcription is shown as number of reads overlapping each nucleotide. Imprinting scores are shown where at SNPs with sufficient coverage to enable quantification of allelic expression (see Materials and Methods). (B) Imprinting confirmation of *Grb10as* by cDNA-RFLP. (C) Confirmation of imprinting by Sanger-sequencing RT-PCR products. Top trace of each box corresponds to embryos with C57BL/6J mothers. Base at position 3 is polymorphic. In all cases, bias agrees with sequencing data on paternal influence.

Even in the extensively studied Prader-Willi-Angelman syndrome locus, we find multiple sites of novel imprinted transcription. Imprinted sequencing reads and contiguous transcription up to 3 kb downstream of Frat3 provide evidence for a Frat3 (Peg12) 3'UTR extension. We also observe contiguous, paternally biased transcription extending up to 700 kb from *Mkrn3* on the negative strand, encompassing transcription antisense to *Ndn* and *Magel2*. 
Besides extensive ncRNA transcription, we also identified coding transcripts that had not been previously shown to be imprinted. Within 300 kb of the Mest locus, we observed a maternally expressed gene, mKIAA0265, predicted to encode a protein containing multiple Kelch domains. Adjacent to the paternally expressed Sgce gene of the Peg10 locus, we verified the maternally biased expression of the Casd1 gene, which encodes a glycosyl transferase and is highly expressed in the brain, like other known maternally expressed genes in the cluster such as neurabin and the calcitonin receptor. We also identified three examples of genes that exhibit parent-of-origin sex-bias and that are more than 10 Mb from any other known imprinted genes. All three (Zdbf2, Pde4d, Tbc1d12) exhibited paternal bias (Table 1).

**Discrepancy between these data and report of ~1300 novel imprinted loci**

While these data do suggest that other imprinted transcripts remain to be identified, they suggest that application of a comparable approach will likely lead to discovery of a comparable number of novel imprinted transcripts when distinct tissues or stages of development are analyzed. Since this prediction differed from the report of ~1300 novel imprinted transcripts, we explored the basis for the discrepancy.

**Novel imprinted loci in embryonic brain do not replicate**

To distinguish known from novel imprinted loci we compiled a catalog of genomic coordinates for all 128 known mouse imprinted genes that we were able to recover from the literature [251, 384, 385, 401-406] (accessible from GEO [407] under accession GSE27016). All 128 were published prior to the recent papers reporting many more imprinted loci [386, 387].
E15 brains yielded considerably more novel imprinted genes than either the adult PFC or POA (553 vs 153 and 256 respectively), providing the richest opportunity to test reproducibility [387]. Aside from an inexact match in developmental time points (E17.5 vs E15), our experiment was a faithful reproduction of the approach used by Gregg et al. We both used brains from reciprocally crossed B and C F1s (from here on BxC will be used to describe F1s derived from B mother and C father; CxB will denote the reciprocal). We both constructed sequencing libraries using the standard Illumina RNA-Seq protocol and sequenced them to 36 bp (single-end) on an Illumina platform. We both used Novoalign (www.novocraft.com) to map reads to UCSC mouse transcripts and ncRNAs from the functional RNA database [408]. We used the same set of SNPs [387] and the same criteria for identifying imprinted transcripts (i.e. containing at least one SNP with 10 or more reads with reciprocally biased expression, \( p < 0.05; \) chi-square test). We observed 100% agreement on known imprinted gene calls in E15 brain, POA, and PFC [387], confirming that our analyses were consistent.

We detected 38 and 42 known imprinted genes in E17.5 and E15 data respectively. 32/42 (76.2%) were detected in both samples (0.1 transcripts expected by chance; Figure 7). This was in sharp contrast to 51/396 (12.9%, 24 expected by chance) novel imprinted genes that confirmed in our screen (Figure 7). This discrepancy is not inconsistent with the experimental validation carried out on novel imprinted genes by Gregg et al.: included in these 51 replicating genes were 2/2 with no previous evidence of imprinting (\( Eif2c2 \) and \( DOKist \)) that were discovered and further validated in E15 brain [387].
To investigate the benefits of sequencing parents to identify heterozygous SNPs, we repeated this analysis using publicly available SNPs. Perlegen [390] used microarrays to resequence the CAST genome in 2007 and the Wellcome Trust Sanger Institute sequenced 17 mouse strains and released ~19 million C57Bl/CAST SNPs in 2009 [410]. We converted SNP transcript coordinates published by Gregg et al. [387] to genomic coordinates (July 2007 NCBI 37/mm9) using coordinates of UCSC Known Genes [393]. 99.96% SNPs [387] mapped successfully to 136,532 unique positions. 88.9% of these were among the 19.6 million SNPs identified by Perlegen [390] and/or Sanger [410], of which 98.9% agreed on base identity. The transition:transversion ratio of SNPs that agree with [390, 410] was 3.00, and 2.06 for the remaining 11% novel SNPs, suggesting that the novel SNPs reported by Gregg et al. [387] have a higher proportion of false-positives. False-positive SNPs cannot lead to an imprinting call since there would be no reads supporting the non-reference CAST allele. Reanalysis of the data using only the 88.9% SNPs that also exist in the public domain yielded nearly identical results (Figure

![Figure 7. Independent replication with E17.5 brains.](image-url) Replication recapitulates 76.2% (32/42, 0.1 expected by chance) known and 12.9% (51/396, 24 expected by chance) novel imprinted genes reported previously [409]. 12,171 (E15) and 10,418 (E17.5) genes were sufficiently powered for detection (≥1 SNP with ≥10 reads), of which 10,222 were in both samples. Reanalysis excluding SNPs detected in parental transcriptomes while relying exclusively on public SNPs marginally changes the outcome of imprint detection (numbers shown in brackets).
7; see numbers in parentheses) with no reduction in sensitivity for known imprinted genes and less than 3% sensitivity reduction in novel regions. This demonstrates that sequencing biological parents when SNPs are publicly available from sequenced parental strains provides little added benefit [390, 410].

ASE measured with RNA-Seq is exceedingly more noisy than accounted for by basic counting statistics.

Statistical modeling of allele-specific expression measured by transcriptome sequencing is an unresolved challenge [411, 412]. Gregg et al. [386, 387] used a chi-square metric that assumes no experimental biases are introduced during library construction, sequencing, genomic alignment, and that each sequencing read is independent of all other reads. Unfortunately these assumptions are often violated [412-416], and systematic error in quantifying allele-specific expression by RNA-Seq is just now becoming apparent [411, 417-419].

To begin to understand the underlying cause of inconsistent imprinting calls we investigated the accuracy of ASE quantification with RNA-Seq. It has previously been shown that ASE measured at the same SNP is highly reproducible across technical and biological replicates [387, 420]. However, this comparison is immune to systematic error such as priming, fragmentation, and PCR biases that arise during library construction, sequencing chemistry [418], and read alignment [413]. Since most sources of systematic error are sequence-dependent, a more informative test would compare concordance in ASE within the same sample, but between independently sampled sites where the level of ASE is the same.

Technical and biological variation explain the majority of novel imprinted genes.
For discovery of imprinted genes, a simple negative control that accounts for systematic error, technical variation, and biological variation is to ask how many SNPs/genes exceed significance in a mock reciprocal cross (i.e. comparing samples with the same parental background as though they were from reciprocal crosses). In such a comparison any reciprocally biased expression cannot be caused by genomic imprinting and is a measure of the technical and biological variation of the experimental approach. Data from two animals of opposite sex was available for PFC and POA and enabled two mock-cross analyses (Figure 8A). Strikingly, in the mock reciprocal cross, nearly as many imprinted gene calls exceeded the significance threshold used by Gregg et al. as in the reciprocal cross (Figure 8B). Similar to our comparison in embryonic brains, the majority of the known imprinted gene calls from reciprocal crosses were the same, but novel calls were not (Figure 8C). Comparing males to females in the reciprocal analysis controlled for sex-specific expression biases. We confirmed that mock-reciprocal hits are not caused by differences in sex by generating additional sequencing data from a male E17.5 brain sample. This enabled a comparison based on true (sex-matched) biological replicates and revealed that male vs male and male vs female BxC mock comparisons produced equivalent numbers of false positive measurements (Figure 8D). While we used approximately half of the data to generate calls (Figure 8B), randomly removing mapped
Figure 8. False discoveries explain the majority of novel imprinted genes. (A) Schematic of experimental design and mock (negative control) comparisons. The red arrows indicate mock comparisons, green arrows indicate same-sex reciprocal comparisons, while blue arrows indicate mixed-sex reciprocal comparisons. (B) Number of SNPs/genes that exceed threshold used by [386, 387] in reciprocally crossed and mock PFC samples. (C) Agreement on known and novel imprinted genes detected in the two reciprocal comparisons diagrammed in Fig 8A. (D) Number
reads from aggregated data revealed that sensitivity is not markedly different at 50% vs 100% of input reads (Figure 8E). Furthermore, the estimated proportion of novel imprinted genes that are false-positives is not impacted by further down-sampling (the slope of the line is consistent when sufficient data exists to overcome noise; Figure 8F) and an overestimate on account of reduced sequencing depth is thus unlikely. We note that an aggregate mock comparison (1+4 vs 2+3) is not an informative negative control since potential sex-specific imprinted genes would not be balanced in this scheme and the output would be a mixture of true sex-specific imprinted genes and false-positives where the contribution of each is not clear.

A high false-discovery rate may also explain the large number of sex-specific imprinted genes reported by Gregg et al. [386] since these, by definition [386], only reach significance in a comparison of one reciprocal cross (e.g. between males) and not the other (e.g. between females). If this were a reliable assay for identifying sex-specific imprinted genes then nothing should meet significance by comparing opposite sexes within each cross, since expression at a sex-specifically imprinted locus would always be biallelic in one animal. 51 imprinted genes in reciprocally crossed male PFC samples reached significance that did not reach significance in reciprocally crossed females at the threshold used by Gregg et al. (but had sufficient coverage to make a call; 36 agreed with Gregg et al. [386], \( p<1\text{e}-63 \)). However, a similar number reached
significance in negative controls (63 genes in mixed-sex and 39 in mock reciprocals; Figure 9). We obtained similar ratios for female-specific PFC imprinted genes and all POA comparisons (Figure 10), demonstrating that this approach is not sufficiently powered at the selected threshold of statistical significance.

**Figure 9. False-discoveries explain the majority of sex-specific imprinted genes.** Frequency of SNPs meeting the criteria used in [386, 387] to report sex-specific imprinted genes when comparing male PFCs from reciprocal crosses (A) as in [386, 387] as well as mixed-sex reciprocal (B) and mixed-sex mock comparisons (C). Red points (○) are SNPs that are biased toward the same parent-of-origin (exceed 10 read counts, \( p < 0.05 \)) in the comparison indicated on the x-axis but not in the comparison indicated on the y-axis (the criteria in [386, 387] for sex-specific imprinted gene calls). Blue points are SNPs exceeding significance in comparison of animals on both axes.
Figure 10. The majority of sex-specific imprints reported [387] are due to technical and biological variance. Red points (o) are SNPs that are biased toward the same parent-of-origin (exceed 10 read counts, $p<0.05$) in the comparison indicated on the x-axis but not in the comparison indicated on the y-axis (the criteria in [386, 387] for sex-specific imprint calls). Blue points are SNPs exceeding significance in comparison of animals on both axes.
Discovery of novel imprinted genes

To estimate the total number of imprinted genes we first asked how many are detected in the four available datasets (E15 brain, E17.5 brain, adult PFC and POA). Aggregating allele-specific reads across SNPs in the same gene improved our sensitivity in known imprinted regions (data not shown) and we thus applied this approach genome-wide. We also took advantage of all publicly available SNPs [387, 390, 410] and expanded our alignment reference to include the whole genome (see Materials and Methods). Using the mock/reciprocal approach to estimate false-discovery (Figure 11A), we proceeded with \( p = 1 \times 10^{-4} \) (FDR<0.05) as a threshold of significance for calling a gene imprinted (in addition to the standard reciprocal bias toward sex of parent and \( \geq 10 \) reads in each cross). We selected \( p = 1 \times 10^{-4} \) as a threshold (e.g. as opposed to \( 1 \times 10^{-3} \)) since candidates with scores between \( 1 \times 10^{-3} \) and \( 1 \times 10^{-4} \) did not validate by pyro-sequencing (see below) and sensitivity is negligibly impacted (Figure 11A,B). We identified a total of 53 putative imprinted genes in at least one sample (Figure 12). 5/53 occurred in all 4 samples and 3 (\textit{Eif2c2}, \textit{Cdh15}, and \textit{DOKist4}) were validated by Gregg et al [387]. We also detected 56 genes that were previously known to be imprinted (27 in all 4 samples). Of the putative novel genes, 4 are probable extensions of known imprinted genes based on EST or transcription evidence derived from this data (3 of the 5 putative novel imprinted genes which recur in all 4 samples), 7 others are associated with known imprinted clusters (within 1 Mb) and 42 are completely novel (Figure 12).
Figure 11. Validating features predictive of genomic imprinting. (A) Estimated false-discovery rate as a function of \( p \)-value threshold. Allele-specific expression counts were summed over all SNPs within each gene. False-discovery was estimated as the number of significant genes in a mock cross (1 vs 3 + 2 vs 4; see Figure 8A divided by the number of significant genes in a reciprocal cross (1 vs 4 + 2 vs 3). Sensitivity was computed as the number of known imprinted genes meeting significance in the reciprocal cross (1 vs 4 + 2 vs 3) as a fraction of imprinted genes powered for detection (at least one SNP with \( \geq 10 \) reads in both animals). (B) Imprinting scores (left panel), allelic bias (middle panel), our imprinting call and pyro-sequencing call (right panel) for 37 loci tested by pyro-sequencing. Imprinting Scores are \( \log_{10}(p) \), where \( p \) is the less significant \( p \)-value from chi-square tests performed on the two reciprocal crosses; negatives arbitrarily represent paternal bias. Expression bias is the allelic bias measured in RNA-Seq data from the average of embryonic brain data (E15/E17.5). Imprinting call is based on an imprinting score threshold of 4 (\( p \leq 1e-4 \)) in E15 and/or E17.5 brains. Pyro-sequencing validation was done on total RNA from E17.5 brains and calibrated on biological replicates (see Materials and Methods for details). Imprinted Loci reported by Gregg et al. represent 17 randomly selected SNPs reported imprinted in E15 [387].
Figure 12. Genomic clustering of candidate imprinted transcripts. Imprinting scores of candidate imprinted genes discovered in E17.5 brain data and reanalysis of E15, PFC, and POA samples [387], clustered by genomic coordinates.

Pyro-sequencing validation of novel imprinted genes in E17.5 brain

From manual inspection we identified three distinguishing features among the known imprinted genes that we detected and reasoned these may be useful for predicting novel
imprinted regions. These include: 1) reciprocal allelic bias and high sequencing depth (reflected in conjunction as the imprinting score), 2) agreement on imprinted expression among neighbouring SNPs, and 3) recurrence of signal across biological replicates and/or tissues. To further investigate the predictive potential of these features we identified 37 candidate imprinted loci that represent a range of values for each feature and tested these by pyro-sequencing. In addition to 5 positive controls, we tested 17 candidate loci selected at random from the list of imprinted SNPs reported by Gregg et al [387] in E15, 4 candidates with marginal imprinting scores (between 1e-3 and 1e-4), 2 candidates detected in adult but not embryonic brain, 7 candidates detected only in embryonic brain, and 5 candidates detected in at least two samples. All 17 candidates reported Gregg et al. are from the 'complex' category where the imprint does not agree with other SNPs in the same gene (this category accounts for 94.8% of the novel imprinted genes reported [387]).

Pyro-sequencing validation suggests that all three features are predictive (Figure 11B; Figure 13). To establish a level of technical and biological variance in our pyro-sequencing assays, we first measured ASE for the 37 loci in biological replicates (two BxC E17.5 brain samples) and observed excellent agreement (Figure 1). This also enabled us to establish a meaningful threshold for detecting differences in ASE (see Materials and Methods). In agreement with our results suggesting that the majority of the Gregg et al imprinting calls are false-positives as well our own imprinting calls on these 17 loci (16/17 predicted to be negative), none validated by pyro-sequencing (Figure 11B). Of 11/16 predicted parent-of-origin effects in embryonic brain (incl. 5 positive controls) that validated, all contain more than one SNP where parental bias was observed, suggesting that consensus SNP calls may be a valuable predictor. The imprinting score was also predictive; the average (absolute)
Figure 13. Pyro-sequencing traces confirmed 6 of the 11 predicted novel parent-of-origin effects in E17.5 brains. For each gene (A-L), (i) are traces from E17.5 brains from BxC parents, (ii) are traces of E17.5 brains from CxB parents and (iii) are traces on genomic DNA from BxC parents. (A) Frat3 is a positive control. (B-G) Confirmed novel parent-of-origin effects. The Bcl2l1 traces illustrate the utility of sequencing 50:50 hybrid BxC DNA to calibrate each assay. In this particular assay an unknown effect caused a slight allelic bias (57:43) even in the case where input was 50:50 B:C DNA. If assay bias is not accounted for, the Bcl2l1 reciprocal F1 traces suggest a genetic background effect on allelic bias, but support a parent-of-origin effect once assay bias is accounted for by normalization to the 50:50 hybrid DNA. Please see Materials and Methods for details on estimating significance of bias from technical replicate runs. (H-L) Predicted novel parent-of-origin effects that did not validate. In all cases, SNPs are highlighted in yellow with the percent contribution of each nucleotide superimposed on each. Note that the
sequencing chemistry is such that adenosine produces 16% more signal than other nucleotides in the trace (Qiagen, personal communication) and the Pyromark software accounts for this when calculating the allelic ratio.

value of the 6/11 novel parent-of-origin effects that validated was 28.4 vs 6.7 for the 5/11 predictions that did not validate and none of the four tested predictions in the 3-4 range validated (Figure 11B). 5/7 predictions detected in only one of the embryonic samples as well as the two detected only in POA did not validate, suggesting that recurring detection in more than one related sample is also informative.

Discussion

Imprint discovery using parallel sequencing

We have used high throughput sequencing to carry out a conceptually simple, genome-wide screen for imprinting in mice that were physiologically normal. Using strand-specific total RNA amplification provides an advantage over common methods that rely upon polyA priming, and allows the detection of expressed SNPs in entire transcripts, including intronic regions and non-polyadenylated transcripts. The advantages of sequencing over microarray-based assays include experimental design without a priori knowledge of SNP position or transcript sequence, and digital readout of transcript abundance, which drastically reduces uncertainty associated with microarray probe cross-hybridization. The success of this approach is reflected in our ability to detect 14/17 known imprinted mouse loci (6/6 classical regions [421]) that have been identified over the last 25 years.
Nearly all of the known imprinted transcripts that were not detected here either did not contain detectable polymorphisms or were not adequately expressed in the embryo. Of 90 known imprinted transcripts [381, 422], 36 contained SNPs and were sufficiently expressed to make a call on expression bias. Of these 36, 32 (89%) had the expected parent-of-origin bias at a confidence of p<0.01 in one of the two crosses, and 27 (75%) exceeded p<0.01 confidence in both crosses. Increased sequencing reads and additional resolution by monitoring more SNPs from crosses in additional strains, such as the divergent PWD/PhJ or MOLF/EiJ strains [390], will improve sensitivity in an enhanced assay of E9.5 embryos.

This study also extended the number and characterization of imprinted coding genes. mKIAA0265 and Casd1 are novel imprinted genes within the Mest and Peg10 locus, respectively [423, 424]. Intriguingly, we observed paternally biased expression of several individual genes, far from any known imprinted loci. On the other hand, our study did not detect any new imprinted loci comparable in size or expression bias to such well-known loci as Igf2r and Dlk1. Although future efforts will increase the sensitivity of our screen, it is also possible that the imprinted loci that have yet to be discovered will have different properties from classic loci. For instance, they may have distinct tissue-specificities or levels of allelic bias.

How many imprinted genes are there?

A further conclusion of this work is that there is no evidence for mammals having an order of magnitude more imprinted loci than was previously appreciated, as two recent papers claim [386, 387]. Independent replication of this work, reanalysis that included negative controls to estimate false-discovery, and follow-up validation using an independent assay unilaterally suggest that the vast majority of the reported imprinted genes are false-positives explained by variation in the assay and experimental approach. This is in agreement with long-standing
genetic estimates and our earlier screen using parallel sequencing, which is effectively the same global approach.

Is it possible then that hundreds to thousands of undiscovered imprinted genes exist? Yes, but there is no evidence for it. These would need to occur in developmental stages or tissues not yet assayed or represent transcripts that are invisible to standard RNA-Seq. Examples of imprinted expression that evade RNA-Seq are non-polyadenylated transcripts, antisense transcripts imprinted/expressed to similar degrees and in opposite directions such that the signal cancels out, and biases that occur to such a minor extent that they are detected below the threshold of noise. Because RNA-Seq cannot be modeled with counting statistics that assume each read is randomly and independently sampled and free of systematic bias, it is imperative that this threshold is firmly defined. Without accounting for background, SNPs with very small parent-of-origin biases resulting from assay variance may appear imprinted with high statistical significance, particularly if the SNP is highly expressed. Even if some of these were real, our pyro-sequencing efforts did not validate any. One could still argue that the effect is below the threshold of pyro-sequencing detection (e.g. less than a 5% difference in this study), but if this were the case then the effect on transcriptional load would be small and likely without functional consequence. In any case, the current state of RNA-Seq and analysis cannot detect these minor effects, even if they do exist.

A more fruitful application of RNA-Seq to imprinting discovery in the near term may involve screening additional developmental time-points, tissues, and species. RNA-Seq methods that do not require polyA+ selection and that retain strand-specificity may also uncover novel transcripts of the noncoding/antisense variety, some of which have already been shown to be important for establishing and maintaining imprinting states [425]. Finally, integration with
complementary global datasets, such as genome-wide allele-specific methylation maps, will improve both specificity of imprinting discovery as well as insight into the underlying mechanisms.

Overall, 6/11 novel parent-of-origin effects validated by pyro-sequencing: U80893, Ccdc40, Bcl2l1, Mapt, Adam23 and Wars. Only Adam23—where mutants are smaller than wild type littermates—appears to have a growth phenotype [426], although this is confounded by their tremors and ataxia [426]. The function of several others is known: Ccdc40 is essential for left-right patterning in both mice and humans, full range cilia motility and a causal mutation for a variant of primary ciliary dyskinesia in humans [427]; Bcl2l1 inhibits apoptosis [428]; while Mapt stabilizes and helps organize microtubules within neurons [429]. Three of these genes, Wars, Ccdc40, and Mapt, were marginally biased in their parent-of-origin expression (Figure 11B and Figure 13). This may be due to a mechanism that causes incomplete silencing or due to allele-specific expression in only a subset of cell types/tissues comprising the organ assayed.

Using more homogeneous samples provides an obvious path forward, but revealing the functional consequences to these types of imprinting cases is the real challenge since many novel imprinted genes awaiting discovery will likely fall into this category.

An estimate for the total number of imprinted genes must account for several variables, including the increasing difficulty in validating novel candidates. Of the 6/11 candidates that validated in this study, only 2 of the 6 represent parent-of-origin effects that are clearly not associated with previously known imprinted regions (Figure 11B and Figure 12). This severely limits our power to establish a firm number and an estimate should be interpreted with caution. Nonetheless, assuming a confirmation rate of 37.5% (3/8) for the remaining untested embryonic brain candidates and a detection sensitivity of 56/128 (known imprinted genes) extrapolates to an
estimate of 37 imprinted genes awaiting discovery and validation, yielding an estimate for a total number of \( \sim 175 \) imprinted genes.

Lessons learned: implications for future imprinting discovery with RNA-Seq

Although the statistically-derived imprinting score is the strongest predictor of imprinting, it cannot be interpreted as a direct measure of probability. As others have noted, RNA-Seq is not free of systematic error \([411, 413, 417, 418]\) which may impact measurement of ASE. That said, since detection of imprinting requires demonstration of reciprocal bias, the effect introduced by systematic error (which is sequence-dependent) can be partially mitigated by estimating FDR from a mock-reciprocal analysis (Figure 11A).

Second, recurrence of imprinted expression across related samples increases the likelihood that novel candidates validate. Although exceptions exist \([251]\), if expression of a given gene is imprinted in adulthood, its expression is generally also imprinted in the developmental precursor of that tissue. Data in Wamidex \([251]\), for example, suggests \( \sim 90\% \) concordance between imprinted expression of an adult tissue with its precursor tissue. 20.0\% and 75.0\% of the putative novel imprinted genes in the POA and PFC respectively were also imprinted in the E15 and/or E17.5 brain (Figure 14). These numbers increase to 68.5\% and 93.3\% respectively, if known imprinted genes are
Figure 14. Concordance in regional imprinting for 53 novel imprinted genes at significance exceeding $p<0.0001$. 
included. The rate of independent confirmation amongst putative imprinted transcripts was higher for those reaching the threshold in multiple samples, suggesting that these figures are likely an underestimate.

Third, consensus in parent-of-origin allelic bias among neighbouring SNPs provides additional predictive potential. Read aggregation across SNPs in the same gene enabled detection of concordance (the same gene exhibiting imprinted expression in more than one sample) in several genes that appeared biallelic using \( p < 0.05 \) without read aggregation [387]. Examples include \textit{Wars} (POA), \textit{Adam23} (POA), \textit{Klhdc10} (E15, POA), and \textit{Cdh15} (E15, PFC).

Fourth, proximity of novel candidates to known imprinted regions or to each other is also predictive. Most known imprinted genes occur in clusters that can span hundreds of kilobases [384] and typically share regulatory mechanisms. 7 (of 8) putative novel imprinted genes that were identified in at least two samples are associated with known imprinted regions (<1 Mb), as well as three detected in only one sample (Figure 12).

In conclusion, until we can accurately model ASE measured with RNA-Seq, estimating FDR of imprinted gene discovery will ideally be done empirically. The additional criteria mentioned above can be used to further rank novel candidates, but since no combination of criteria was absolutely predictive among the novel imprinted genes identified in this study, we assert that independent validation is essential for making definitive claims about imprinting of any gene.
General Discussion
The molecular network endowing a cell with the capacity to differentiate into cell types from all three germ layers is a cornerstone of developmental biology. Before mechanistic insight provided by molecular biology, differentiation was modeled as a canalization process where the differentiation options of pluripotent cells continually narrowed until highly specialized cell types with limited potency are established [430]. Although many details remain to be elucidated, molecular data has provided a framework for a mechanistic understanding of these processes. Lineage differentiation is driven by feed-forward and feed-back loops of TFs and ncRNA that culminate in stable expression of regulons. These regulons collectively comprise the functional elements of a given cell type. Canalization, or progressive restriction of differentiation options, is established by the absence of TFs required to activate certain regulons as well as epigenetic modifications rendering certain genomic loci inaccessible to activation.

The molecular mechanism of pluripotency
How pluripotency is established and maintained remain unclear at the molecular level. Without a complete molecular understanding pluripotency can still be defined functionally: competence of a clonally-derived population to generate cell types from all three germ layers, as well as by factors that mark or promote this state. Intrinsic factors, such as Oct4 and Nanog, are commonly used as markers of pluripotency, but it is unclear whether the inability of every cell in a population homogenous for pluripotency markers to meet functional criteria of pluripotency relates to technical issues or heterogeneity in the population [136]. Current models cannot provide a definitive pluripotency signature, suggesting that essential considerations remain unknown.

Pluripotency is described as a stable state in which a core network of factors auto-regulate and buffer against lineage induction by extrinsic signals [113, 131]. Alternatively,
pluripotency factors are modeled as lineage specific inducers that are each simultaneously promoting maturation into a lineage and blocking others [135]. In the latter scenario, stability is achieved by auto-regulatory activity as well as activation and repression of separate lineages by each pluripotency factor such that equilibrium is established by different pluripotency factors simultaneously activating and repressing each lineage. Differentiation occurs when the equilibrium is disrupted by a pluripotency factor exceeding the buffering capacity and activating a lineage-specific program [135].

In accordance with the second model, pluripotency factors are commonly interpreted as lineage-specific activators since manipulation of their transcript abundance increases the prevalence of certain cell fates in culture. If Oct4 abundance is too high, differentiation into primitive endoderm and mesoderm, cardiomyocytes or neurons occurs depending on culture conditions [143, 144, 267]; if it is too low, TE is produced [267]. Oct4 depletion in ES cells mirrors in vivo results, where TE differentiation is thought to be a binary decision. It occurs when Cdx2 levels increase relative to Oct4. They have a mutually antagonistic relationship that amplifies differences between the two, and promotes lineage segregation wherein Cdx2+ cells result in TE [431]. In vitro and in vivo data concerning the consequence of increased Oct4 do not agree as well. In vitro reports propose that Oct4 promotes mesoderm and represses neural differentiation. However Oct4 is present through the specification of both mesoderm and neuroepithelium in embryos, persisting in both lineages for ~2 days after specification [142]. This may indicate that the role of Oct4 in specification of separate lineages is distinct, being modulated by other factors. One alternative interpretation for this discordance is that over-expression of Oct4 is producing results in vivo that are not representative of its embryonic role. If Oct4 over-expression enhanced survival, proliferation or repressed alternative stochastic lineage
differentiation in the culture conditions used, this would also give the perception of enhanced lineage differentiation. Since the cell fates observed following over-expression are consequences of transcript abundance that likely exceeds the levels driven by endogenous Oct4 regulatory elements, they may not representative of Oct4’s role in development.

Alternatives to a levels based model

Oct4 over-expression enhancing lineage differentiation does not demonstrate that it is a lineage-specifier. Oct4 has been implicated in regulation of each of the alternatives described: survival, lineage repression and proliferation. For example, depletion of Oct4 in PGCs leads to apoptosis [253], while over-expression induces hyperplasia and prevents terminal differentiation in the epithelium of the intestine [271]. These alternatives confound interpretation of assays that suggest lineage induction based on lineage bias. The majority of murine cell types do not require Oct4 for viability, it is insufficient for proliferation based on the G1-arrested ICM during diapause which is Oct4+ [432, 433], and its presence does not block certain differentiation programs [143]. These results indicate that Oct4 either has context-specific effects or that over-expression produces ectopic effects. A more specific example clarifies this point. ES self-renewal is co-dependent on Nanog and Oct4 [325]. Nanog and Stat3 are present in limited quantity in ES cells and inhibit differentiation to primitive endoderm and mesoderm [325]. As such, Oct4 over-production may titrate Nanog and relieve repression of lineage entry, as opposed to actively inducing differentiation of primitive endoderm and mesoderm itself. The first model of pluripotency would explain this scenario by suggesting that Oct4 is not inducing differentiation, but by titrating Nanog and Stat3 it enables another signal that did not meet the threshold to induce differentiation when stoichiometry of the pluripotency network was sufficiently balanced to mitigate it. The second model would argue that when stoichiometry is
disrupted, Oct4 activates lineage-specific factors in the primitive endoderm and mesoderm programs. Disruption of Oct4’s DNA binding by manipulation of the homeodomain and POU domain such that it could titrate Nanog but not activate transcription would resolve whether it is essential for activating these lineage differentiation programs.

Our data suggest that Oct4 is not solely a lineage-specifier. Regulation of Sox2 by Tbx6 determines mesoderm versus neural fate in tailbud progenitors, and provides one illustration of the phenotype anticipated from disrupting a lineage-specifier [434]. In the absence of Tbx6 ectopic neural tubes form at the expense of paraxial mesoderm [313], while during formation of paraxial mesoderm Tbx6 represses Sox2 [434]. This example illustrates cell fate conversion accompanied by continuous proliferation, indicating that a diminished cell-cycle is not an inherent requirement of cell-fate alteration. Our data are fundamentally different. Proliferation of posterior progenitors decreases dramatically when Oct4 is depleted, indicating that Oct4 is either not a lineage-specifier or mediates both cell fate and proliferation.

Resolution of the alternative pluripotency models

Ultimately, the distinguishing feature between the two models is that pluripotency factors activate lineage differentiation in the second. They are conceptually distinct in that the latter model implies that Oct4 is part of a feed-forward loop or other intrinsic means of determination, while the former suggests that extrinsic factors such as morphogen gradients or short-range signals eventually over-ride buffering of Oct4 and the pluripotency network. Details of whether altered survival, proliferation and/or repression of specific lineages contribute to enhanced lineage differentiation will provide clarity. Clarification of the genomic sites where each pluripotency factor is activating or repressing transcription are needed to distinguish these models. At present the code dictating whether sites bound by Oct4 are activated or repressed is
not understood. Correlations between the number of TFs and transcriptional activation have been reported, but do not offer robust predictive capacity. Disruption of Oct4 binding and assays of the associated transcriptional response for a large panel of individual loci will inform genomic approaches such that a predictive relationship between TF binding and transcriptional response can be definitively established.

**Understanding Oct4’s mechanistic contribution to pluripotency**

The essential molecular contribution Oct4 makes to pluripotency is unknown. By way of contrast: Nanog obviates the requirement of Lif/Stat3 signalling, indicating that it is their essential target. Similarly, constitutive Oct4 production obviates the need for Sox2, suggesting Oct4 is the only required target of Sox2. No such convergence point downstream of Oct4 has been established. Oct4’s contribution is generally modeled as being transcriptional regulation, because it contains DNA binding domains and is an established TF at certain loci, but it is not known which subset of targets are relevant to pluripotency. Alternatively, protein-protein interactions suggest that it may have other functions in the nucleus, primarily chromatin remodelling, but also in recombination and repair, DNA replication, nuclear organization as well as other miscellaneous functions [121]. Whether these interactions are essential for pluripotency is unclear, but might be resolved by varied mutagenesis disrupting different protein-protein actions to dissociate the diverse array of proposed functions that are thought to be interaction dependent.

One broad alternative to lineage repression is that Oct4 contributes to pluripotency by promoting a unique cell-cycle. G1-arrest of the ICM of embryos in diapause presents a clear distinction with ES cells that do not have a G1/S checkpoint [132]. The presence of Oct4 in such G1-arrested ICMs indicates that it is insufficient to promote the cell-cycle [435], however its
removal from mitotic cells correlates with diminished proliferation. Reduced proliferation upon Oct4 depletion is commonly interpreted as differentiation due to loss of potency rather than cell-cycle arrest. Distinguishing whether loss of proliferation precedes differentiation or the other way around might be accomplished by assaying whether Oct4 depletion results in the same cell fate being specified when cells are encouraged to cycle by other means, perhaps driving the cell-cycle with ectopic expression of cyclins. Alternatively, if the cell-cycle was not disrupted when Oct4 is depleted and downstream lineage-specifiers are repressed, this would suggest that changes in the cell-cycle are secondary to lineage repression. Using the case of TE differentiation as an example, if Oct4 is depleted in an ES line deficient for requisite mediators of TE differentiation (perhaps Cdx2 and Eomes would be sufficient) and the cell-cycle was not disrupted, this would affirm Oct4’s role as a lineage repressor. Oct4 does inhibit some factors in the TE differentiation program, but TE differentiation in the absence of Oct4 could be reinterpreted as impaired cytokinesis depending on the outcome of these experiments. Overexpression of Oct4 in intestinal epithelium that prevents terminal differentiation also might support such a model where the contribution of Oct4 to pluripotency is driving the cell-cycle and limiting the window during which cells can respond to differentiation signals. Oct4 interacts with ncRNAs that inhibit inhibitors of the G1/S checkpoint in mouse ES cells [132], providing a possible mechanism for how Oct4 might promote the cell-cycle.

Cellular potency in the context of development

Our data illustrate a conserved requirement for Oct4 in posterior expansion, raising the question of whether this role or lineage activation/repression in pluripotent populations found earlier in development is ancestral. Pou5f1 expression is maintained by a proximal enhancer (PE) in the epiblast, and a distal enhancer (DE) in morula, the ICM and PGCs [436], suggesting that
its expression domain might have expanded during evolution. The inference being that one of these two functions may have arisen by fixation of the second enhancer that led to Oct4 being produced in a context with a distinct combination of other factors. The function of Oct4 itself would not necessarily be altered in this scenario.

Comparison with a paralogue offers insight into whether Oct4’s function is the same in early pluripotent populations and the posterior. When the *X. laevis* paralogue *Xlpou9l* is disrupted with morpholinos at the 2-cell stage, *X. laevis* development proceeds, but posterior expansion is impaired. This suggests either that sufficient maternal deposit *Xlpou9l* is present to maintain early pluripotent populations in *X. laevis*, or that *Xlpou9l* was maintained for its essential role in posterior expansion. In either case, the ability of *Xlpou9l* to substitute for *Pou5f1* in maintenance of pluripotent mouse ES cells suggests that the function of Oct4, and at least a subset of its homologs, is the same in posterior progenitors as it is in the ICM. In support of this, our expression analysis shows that Oct4 targets exhibit strong conservation between ES cells and E8.5-E9.0 embryos. I infer that the cellular outcomes differ when Oct4 is depleted from the ICM versus E7.5 embryos because of contextual differences. For example, despite common expression in both the ICM and gastrulating embryos, depletion of Sox2 and Oct4 do not phenocopy as they do when removed from ES cells [124, 267]. This indicates that the regulatory relationship has changed: Sox2 is no longer required for Oct4 production as it is in ES cells and targets a distinct regulon based on the features of Sox2 depleted embryos that do not occur when Oct4 is removed [124]. It is not clear why this is the case: an effect might be masked by presence of a redundant factor, or Sox2 may not activate *Pou5f1* at this stage. It is clear from the differences in phenotypes that their regulons do not overlap entirely. These data support the
function of Oct4 being the same in pluripotent cells and during posterior expansion, but do not clarify whether the ancestral role of Oct4 is in posterior expansion or pluripotency.

Our data revealed that Oct4’s role in expansion of posterior progenitors is conserved with homologues [265], raising the question of whether this or Oct4’s role in pluripotency is ancestral. Since *Pou2* and *XlPou9l* are paralogues in *D. rerio* and *X. laevis* respectively, diminished progenitor expansion during development of these preformation-employing species could be interpreted as cooption from the ancestral role in pluripotency. It is assumed that functional conservation with the axoltotyl orthologue implies conservation of the pluripotency network [224], however the axoltotyl orthologue, *AxOct4*, may also have the same role in posterior expansion. This study has not been performed. To resolve which role—pluripotency or expansion of posterior progenitors—is ancestral, additional phylogenetics and functional studies in basal epigenesis-employing ancestors are required. A homologue of Oct4 has recently been identified in the neoblasts of planaria [437], *Smed-POU-P1*, but its function has not been characterized. It did not turn up in a genetic screen for pluripotency factors in planaria [438], but it is not clear whether this result relates to biology or methodology. A global defect in regeneration following *Smed-POU-P1* disruption would support the ancestral role being in pluripotency, while a posterior defect would support posterior expansion being ancestral. Notably, a homologue for *Nanog* was absent [437]. This was interpreted as meaning another factor played the same role of Nanog in pluripotency of planaria, but may indicate that the pluripotency network is derived, perhaps from the bipotent tailbud population described earlier in relation to Sox2 and Tbx6 that contributes to posterior expansion.

Functional data is available from basal cnidarians which speaks to the issue of whether a role in pluripotency or posterior expansion is ancestral. Polynem is a POU domain protein.
recognized by a polyclonal antibody raised against human Oct4 that also exhibits a relatively conserved expression pattern with Oct4: it is expressed in early embryos before restriction to interstitial cells (i-cells) which are either multi or totipotent SCs [439]. Ectopic Polynem in cnidarian causes induction of cnidarian SC markers (Nanos, Vasa, Piwi and Myc) as well as epithelial neoplasms reminiscent of the effect Oct4 over-expression has in the murine gut epithelium [440]. This observation was used to support Oct4’s ancestral function being in establishment and maintenance of pluripotency [440], however it is subject to the same issue described previously in that transcript levels exceed production by the endogenous regulatory elements and may be ectopic effects.

RNAi knockdown of Polynem provides more useful insight. Knockdown in early animals results in lethality, while use later reportedly increases the number of nematocytes (stinging cells of the neurosensory lineage) [440]. Superficially, this resembles the ES-derived model of Oct4 repressing neural tissue formation during germ layer selection [268]. Considering this result in the context of a cell-cycle based model, knockdown of Polynem in cnidaria or Oct4 in mice might lead to cell-cycle exit and neural differentiation because it is the default pathway at these respective stages [324]. The distinction being that lineage differentiation is prevented by a brief G1/S checkpoint as opposed to repression of lineage-specifiers. Considering this possibility further, Nanog is absent in planaria [438], but the putative Nanog homologue in the preformation-employing medaka, a teleost, is reported to be a cell cycle regulator [441]. This is interpreted as cooption, but in this alternative view the function of Nanog might have been retained, but its expression coopted to coincide with Oct4 and jointly mediate a cell-cycle profile that limits differentiation. In either event, a closer look indicates that the result of increased nematocytes following Polynem knockdown was not normalized and/or quantified. Nor was a
lineage trace of i-cells producing the additional nematocytes in this context performed. This begs the question of whether increased proliferation of the entire hydra (model cnidaria) or selectively amongst nematocytes accompanied loss of Polynem function. If additional experimental rigour confirms that differentiation of i-cells to nematocytes or other terminal cell types is induced via loss of lineage-specifier repression upstream of effects on the cell cycle, this would be compelling evidence that the ancestral role of Oct4 is maintenance of pluripotency via lineage-repression. Measurement of proliferation and cell cycle changes after Polynem knockdown as well as localization of cellular effects with respect to body-axes should distinguish these two possibilities. This will be insightful in establishing which constraints acted during evolution of the pluripotency network, and by extension which aspects of model organism biology will be informative in human cells.

Evolutionary differences that might be expected between mice and humans

Phylogeny and germ line development by epigenesis in both mice and men suggest that the pluripotency network and early lineage differentiation programs should be conserved. Varied analysis suggests that the core components are conserved, but other aspects such as downstream targets are not [442-444]. It is not clear whether this relates to high false positive rates in the data used to derive these networks or a genuine lack of conservation. Functional genetic approaches are needed to distinguish between these two possibilities. There are relevant differences during this window of development that are likely to contribute. For example, equivalent pluripotent cells are thought to be somewhat staggered in mice and men, with human epiblast SC being equivalent to murine ES cells [445]. These differences might also pertain to the rodent embryonic morphology which is unique in its cup-shape at this stage [224, 446]. Gastrulation differs somewhat between the two species as well in that human embryos do not ‘turn,’
suggesting that related aspects of the phenotype may not occur or would be less pronounced in humans. In particular, the effect on NTC might be exacerbated in mice due to dependence on turning.

**Specification**

Lineage maturation integrates external signals with intrinsic feed-forward loops. Our surfaceome analysis revealed proteins that contribute to NSC maturation and function. These cell-surface profiles provide a framework for understanding how cellular responsiveness to extrinsic signals changes during NSC maturation.

**Relevance of the proteins implicated in NSC function**

CSC analysis provided a host of candidates and use of small molecules and siRNA in colony forming assays specifically implicated Epha4, Erbb2 and c-Kit in NSC function. EphA4 is a RTK activated in response to Ephrin B3 (Efnb3) that is expressed in the NS and also in mesoderm after E7.75 [447]. Inhibition of EphA4 in ES-derived NSC populations indicates that it is required for viability of the earliest NSC populations. This result was replicated in adult-derived NSCs, confirming that ephrin signalling through EphA4 is required for viability of all murine NSC populations. Indeed, these observations are in agreement with the independent conclusion that Efnb3 signalling through Epha4 promotes cell viability in the adult subependyma, whereas its absence leads to an increased apoptosis [369, 370].

Erbb2 is an RTK initially expressed in the ICM, and activated in response to Nrg-1. Like cyclopamine, Erbb2 signalling blocks colony formation of ES-derived pNSCs as well as adult dNSCs. *In vitro* Nrg-1 is required for proliferation of neural precursors [371], although the specific requirement for Erbb2, as opposed to Erbb4, was not previously shown. As with Erbb2,
the result confirmed with adult-derived NSCs, supporting the relevance of results derived from the model system.

c-Kit is a RTK activated by stem cell factor, whose onset of expression in the neural tube and other lineages occurs by E8.5 [448, 449]. Our data indicate that inhibition of c-Kit increases the frequency of colony-forming pNSCs, but not ES-derived dNSCs or adult dNSCs. This functional outcome is consistent with the finding that c-kit is only upregulated in pNSCs, not dNSCs. One possibility is that c-Kit mediates quiescence of pNSCs, perhaps as regulator of the pNSC pool size. c-Kit’s function in several cell types was established through mutant analysis [449]. That c-Kit mutants have relatively mild neural defects [450], and that its role in other cell types where it is expressed was not apparent from mutants [449] suggests that redundancies may exist in the embryo for a subset of roles performed by c-Kit. Indeed, the roles of single factors are readily accessible in our simplified in vitro system as minimal growth conditions speak to factors that are sufficient, but may be buffered in vivo. Preliminary data suggest that c-Kit serves a comparable function in vivo. pNSCs thought to reconstitute the pool of dNSCs following injury have recently been isolated in adult murine brains (Nadia Sachewsky and Cindi Morshead, unpublished). c-Kit repression also limits colony formation of adult-derived pNSCs (Rachel Leeder, unpublished). These data support the biological relevance of proteins implicated using CSC analysis of our in vitro model. Systematic functional testing of the remaining panel of candidates will contribute to an integrated view of NSC maturation and function.

Utility of this analysis

Systematic functional screening of the panel of markers identified by surfaceome analysis would reveal mediators of NSC biology, facilitate distinction of NSC subtypes, and help resolve the larger issue of the extent to which NSCs contribute to organism behaviour. More extensive
use of small molecules and RNAi to systematically inhibit candidates in pNSC and dNSC colony forming assays would readily isolate cell-surface proteins involved in survival, proliferation or lineage maturation. Follow-up analysis of the variety we performed would distinguish between these roles. A more detailed understanding of the mode of action of these receptors could be achieved by manipulating candidates downstream of the receptors. For example, if an effect on viability is detected, select inhibition of mitochondrial, TNF and Fas-mediated apoptosis would refine downstream mechanisms. In cases where downstream candidates are not obvious, such as lineage maturation through an uncharacterized receptor, a comparison of transcriptional profiles in the presence and absence of receptor blockade should reveal pathways mediating these extrinsically driven aspects of NSC maturation.

Surface barcodes that distinguish NSC populations would facilitate their characterization. ES-derived NSCs more accurately model SE than SGZ-derived NSCs. Both SE and ES-derived NSCs are multipotent and generate a substantial proportion of neurons when differentiated in culture (>10%) [324, 451, 452]. Conversely, SGZ-derived populations rarely yield neurons in culture [451], and appear unipotent when derived from adult animals [451]. Superficial markers of SE-derived NSCs include LeX, CD133, GFAP and Nestin, while CD24, O4, NeuN and S100β mark progenitors or differentiated cells can be used to deplete these populations from culture [214]. The initial screen would be to test markers to see if they enrich or deplete colony forming cells from the bulk population. Either result would be useful in enriching NSCs. NSC purification would facilitate molecular characterization by reducing noise in biochemical and molecular approaches to teasing apart the network of TFs and genomic targets responsible for stemness in NSCs. Such surface markers would also facilitate identification of NSCs in vivo to facilitate understanding of niche-based regulation. Our data provide an extensive panel of
putative markers that can be used to generate a barcode for sorting NSCs, and may provide additional value in distinguishing regional NSC identity [453, 454].

Finally, studies interrogating the function of NSCs in organism behaviour would be aided by more selective means NSC manipulation. Markers specific for NSCs or subpopulations would enable lineage traces and ablations to elucidate which cells are produced in different contexts and the functional consequences or their absence. Functional screens of the candidate panel might also reveal selective means of manipulating proliferation and directing NSC differentiation \textit{in vivo}.

\textbf{Evolutionary differences that might be expected between mice and humans}

For all of their differences mice and men are quite similar when making a high-level comparison of neurogenesis between the two. Neither retains the capacity for large-scale regeneration, with neurogenesis being largely restricted to the SE and SGZ in both. That said, certain differences can reasonably be expected, based in part on structural differences between the rodent and human SE. The differences include a gap that has been reported between the ependyma and strip of astrocytes in the human SE [455]. Further, no chains of migrating Nbs were seen in human brains, raising the question of whether a rostral migratory stream is present [455]. This begs the more important question of whether the function of NSCs in humans and mice is equivalent [456].

\textbf{Imprinting}

Our analysis followed other systematic genomic interrogations for imprinted genes. A phenotypic screen in offspring inheriting uniparental disomies was the first systematic approach, followed by modern genomic approaches such as applying microarrays to comparable crosses [457]. These approaches were complemented by more ‘focused’ approaches including targeted
interrogation of genes adjacent to known imprinted transcripts (efficient due to genomic clustering of imprinted transcripts), as well as assaying orthologues of imprinted genes.

Interrogation bias is evident in these approaches: towards readily apparent abnormalities in phenotype-driven assays, towards transcripts that are highly abundant and heavily biased in microarray based approaches, towards confirmation of known imprints when extrapolating known imprints to other species, towards tissues where imprinted transcription was previously discovered and towards known imprinted clusters.

Our approach distinguishes itself from previous efforts in several regards: we are directly analyzing the transcriptome so it does not require a phenotype for discovery of imprinted transcription; neither genetic nor reproductive manipulations were employed, so genes implicated using this approach are not artefacts of contrived experimental conditions; and sensitivity is considerably improved as RNA-seq data has relatively little noise compared to microarrays. Refinement of the analysis associated with this approach should enable accurate systematic detection and discovery of imprinted transcription.

Systematic detection of imprinted transcripts will enable the complete set of imprinted transcripts to be identified in a range of species, and defined both spatially and temporally. This will facilitate data-driven models of the cause and consequences of imprinting. The conflict hypothesis was primarily appealing because many imprinted genes are growth regulators [458]. Independent evolution of imprinted expression in flowering plants could also be considered validation of this model as another situation where resource allocation is flexible [459]. However, allocation of milk by monotremes is flexible, as is parental investment following delivery in many non-eutharian species. For example, care offered by birds after birth [460]. These behaviours should provide a comparable selective pressure to flexible resource allocation
by the placenta. The conflict hypothesis also struggles to explain variability in the genes which are imprinted in different species, e.g. if the sole selective pressure is growth, why would imprinting of any growth regulators be ‘lost’ in some species after fixation in a common ancestor? This occurred in the case of IGF2R [461], and it seems unlikely that the remaining discordance in genes that are imprinted between species that diverged relatively recently arose solely from de novo imprinting events with no contribution by imprinting loss in a species-specific fashion [251]. Genomic profiling for imprinting in ancestral species will distinguish these two possibilities.

**Improved prediction accuracy will make RNA-seq a sensitive imprint discovery tool**

The putative influences of imprinting –ranging from growth to sexually dimorphic behaviour-- are tantalizing, but in some cases unproven [462]. Imprinted transcription was accurately forecast before supporting experimental evidence was available [241], and hypotheses may also prove accurate when predicting aspects that provide sufficient selective pressure to fix imprinting within a species. Ultimately however, rigorous experimentation is needed to establish the extent of imprinting’s influence.

RNA-seq based methods to imprint identification offer considerable advantage over previous approaches in sensitivity, use of wild type samples, and being phenotype-independent. However, the extent of the technique’s utility is still being determined, and there are limitations. Preparation of genomic libraries can introduce bias; the extent of natural allelic variance is not appreciated; and statistics able to capture reproducible differences require additional empirical validation.

Our analysis indicates three major avenues for mitigating high false positive rates when calling imprints from RNA-seq data. The combination of read depth and strong allelic bias
remain the best predictors of imprinting, but do not capture the complete picture. We found several approaches that help. In the absence of knowledge of the biological and technical drivers of variance in RNA-seq data, mock reciprocal crosses are a more accurate means of estimating false discovery than assuming a normally distributed transcriptome. Lineage relationships are useful indicators, e.g. imprinting in an adult tissue is preceded by imprinting in its precursor, so detection twice within one lineage supports a prediction. Aggregation of statistical confidence across SNPs within the same transcript eliminates the ability to detect isoform specific imprinting events (for which there is precedence, e.g. [463]), but facilitates reduction of the false discovery rate by increasing statistical power. Finally, differentially methylated regions (DMRs) are cis-acting and can affect multiple loci. Our data support the idea that novel imprinted transcripts detected adjacent to known imprinted transcripts are more likely to confirm.

Towards the underlying cause of imprinting: spatial, temporal and phylogenetic studies

The spatial and temporal profiles of imprinting in diverse species may provide insight into the selective pressure(s) dictating imprinted transcription. Imprinting is spatially and temporally specific: genes may be imprinted in a subset of tissues and biallelically expressed in others [251], or imprinted during select temporal windows of development and biallelic otherwise [251]. Moreover, a subset of imprinted transcripts ‘flip’ their expression bias from one parental allele to the other at different stages of development, suggesting context-specific interpretation of the epigenetic markings governing expression [464]. The same allele being imprinted or biallelic in different contexts suggests that imprinted expression is either of no benefit in tissues where expression is biallelic or detrimental. The former would be reflected in a passive imprint loss, where positive selection promotes imprinting in at least one tissue but is absent elsewhere, leading to biallelic expression. This might be reflected by the absence of
tissue-specific enhancers or enzymes required to maintain a given imprint. Conversely if imprinting were detrimental, imprint removal would be reflected in active removal. This might involve tissue specific expression of chromatin remodelers to remove an imprint.

One drawback of the analysis we did is that the RNA-seq data were generated from embryos -- a mixture of tissues. This meant that numerous known and novel cases of imprinting where bias was less than 100% were detected. It is not clear whether these represent stereotypical cases of imprinted expression, 100% from one allele and 0% from the other, which were masked by bi-allelic expression in some tissues or partial inactivation of the imprinted allele. Focusing on individual cell types should mitigate this consideration by avoiding ‘dilution’ of monoallelic expression signals. Novel imprinted transcription may be discovered by analyzing individual tissues: a comparable approach to ours uncovered several novel imprinted transcripts in neonatal mouse brains [385]. These discoveries contribute to one useful parameter by defining the identity of the complete set of imprinted transcripts. Comparison of spatial, temporal and phylogenetic patterns of imprinted expression will provide further data-driven insight. Establishing the extent of conservation in these patterns will dissociate whether spatial and temporal specificity are under tight selective pressure, help infer the identity of this selective pressure based on function of tissues and developmental windows in which genes are imprinted, or simplifying models if conservation is only apparent in a subset of tissues or developmental windows. The latter would support absence of selective pressure outside these windows. The logic being that if a set of transcripts have a common pattern of imprinting their function in this context can be inferred. The value of these atlases will be apparent when coupled with phylogenetic studies where gene identity, as well as spatial and temporal distribution should provide insight into the underlying cause of imprinting.
Towards the underlying cause of imprinting: revisiting function

The literature appears to assume, but not demonstrate, that reduced transcript abundance was the feature of imprinting that was positively selected. This model predicts that a transcript should be lower in abundance when comparing expression of a gene that is imprinted to expression of an orthologue that is not imprinted in a recently diverged species. While this may seem a given on the basis that imprints do reduce expression of one allele, an alternate possibility is that imprinting was fixed during evolution because it limited the expression of one allele in a subset of tissues that are sensitive to deviations or increases in abundance, and provided a fitness benefit to the organism by enabling elevated expression of the same gene in other tissues that are less sensitive to increased abundance. It is implicit in this model that loss of an imprint control region will cause a phenotype when biallelic expression of an imprinted gene occurs in tissues where it is imprinted, the inference is that this imprint is required to facilitate increased expression in other tissues. Thus fixation of a tissue specific DMR in a subset of tissues would have enabled selection of regulatory variants promoting increased expression elsewhere. In effect, rather than imprinting resulting in roughly halving of transcript abundance, the alternative is that it enabled doubling of these transcripts by alterations to the regulatory element in tissues where imprinting does not occur. This alternative model predicts that mice heterozygous for imprinted alleles would exhibit decreased fitness relative to wild-type mice, assuming monoallelic expression in tissues where the gene is not imprinted resulted in halving the abundance of the gene-product. The alternative, passive loss of imprinted expression due to loss of selective pressure in tissues where the genes are normally biallelically expressed, predicts that heterozygous mice would not suffer a fitness defect when competing with wild-type mice.

Comparison in transcript abundance between related species where a gene is imprinted in one
species, but not the other is the most meaningful way to resolve this question. However this is not trivial due to global variation in regulatory elements when comparing species. One approach would be to normalize to a panel of ‘housekeeping genes’ in the transcriptome akin to QPCR methodology. An alternative would be to use a rankings-based method to define transcript abundance in different species and compare whether imprinted genes collectively exhibited lower rankings [465]. More definitive tests would be to switch regulatory elements to ensure that transcript abundance is effected. Would imprinting of a gene in a tissue where it is normally bialleleically expressed cause a fitness defect?

Bystander effects should be integrated into evolutionary models of the causes and consequences of imprinting. Imprinted genes are clustered due to the presence of cis-acting DMRs that can impact the expression of numerous adjacent transcripts. The consequence of this is that if syntenic adjacent to a DMR is lost, the identity of imprinted genes may change. If the relationship between a given DMR and gene is maintained in different species, while the identity of other adjacent imprinted genes is dictated by loss or gain of synteny, this would suggest that selective pressure is positive for only the conserved relationships and that the others may be bystanders. Of course, recombination may yield changes in syntenic that are adaptive in a species-specific fashion, and these two possibilities could be distinguished using genomic insulators to buffer a subset of the imprinted genes in a cluster from the affect of a DMR, or by translocations or transgenics where the impact of biallelic expression of putative bystanders is assayed.

Evolutionary differences that might be expected between mice and humans

Imprinted expression in mammals supports its positive selection, however only about 50% of the genes which are imprinted in mice and men overlap [251]. One possible explanation
for this was discussed in the previous section – that many imprinted genes are bystanders and reflect changes in synteny and drift in the regulatory elements (insulators, spacing, etc) that influence the identity of genes which are imprinted. An alternative is that the majority of species-specific differences are under distinct positive selection in mice and men. The extent of imprinting’s relevance to human biology is unclear. Imprinting undoubtedly sensitizes to genetic mutation, as a single mutation in the expressed allele can result in a phenotype. The related benefits are unknown. Parent-of-origin effects have been detected in mental health conditions and are proposed to mediate complex traits, such as cognitive function [462]. Distinguishing whether imprinting of genes that are not conserved between mice and men were selected as an adaptive trait in mice or result from drift would be instrumental in establishing whether human-specific imprinted transcripts were selected as mediators of complex traits.
References

31. Fraser, H.B. Genome-wide approaches to the study of adaptive gene expression evolution: systematic studies of evolutionary adaptations involving gene expression will allow many fundamental questions in evolutionary biology to be addressed. Bioessays 33, 469-477.


microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell 134, 521-533.


germline reveals similarities with Oct4 and Sox2 and differences between human pluripotent cell lines. Stem Cells 23, 1436-1442.


199

357. Kim, K., Doi, A., Wen, B., Ng, K., Zhao, R., Cahan, P., Kim, J., Aryee, M.J., Ji, H.,
necessary component of cardiac cell compartmentalization during early vertebrate heart
N-cadherin/catenin-mediated morphoregulation of somite formation. Dev Biol 202, 85-
102.
Epigenetic reversion of post-implantation epiblast to pluripotent embryonic stem cells.
Nature 461, 1292-1295.
364. Machold, R., Hayashi, S., Rutlin, M., Muzumdar, M.D., Nery, S., Corbin, J.G., Gritli-
required for progenitor cell maintenance in telencephalic stem cell niches. Neuron 39,
937-950.
Hedgehog signaling by the Hedgehog-binding protein, Hip1, modulates Fgf signaling
are governed by partially overlapping feedback activities of the hedgehog antagonists
368. Stone, D.M., Hynes, M., Armanini, M., Swanson, T.A., Gu, Q., Johnson, R.L., Scott,
369. Furne, C., Ricardo, J., Cabrera, J.R., Pays, L., Bethea, J.R., Mehlen, P., and Liebl, D.J.
(2009). EphrinsB3 is an anti-apoptotic ligand that inhibits the dependence receptor
functions of EphA4 receptors during adult neurogenesis. Biochim Biophys Acta 1793,
231-238.
proliferation and survival in adult neurogenesis. Mol Cell Neurosci 31, 713-722.
371. Calaora, V., Rogister, B., Bismuth, K., Murray, K., Brandt, H., Leprince, P., Marchionni,


specific imprinted gene, TFPI2/Tfpi2, which requires EHMT2 and EED for allelic-silencing. Genome Res 18, 1270-1281.


451. Clarke, L., and van der Kooy, D. The adult mouse dentate gyrus contains populations of committed progenitor cells that are distinct from subependymal zone neural stem cells. Stem Cells 29, 1448-1458.