Epigenomics-Based Identification of Major Cell Identity Regulators within Heterogeneous Cell Populations

Rizwan Rehimi,1 Milos Nikolic,1 Sara Cruz-Molina,1 Christina Tebartz,1 Peter Frommolt,2 Esther Mahabir,1 Mathieu Clément-Ziza,1,2,* and Alvaro Rada-Iglesias1,2,3,*

1Center for Molecular Medicine Cologne (CMMC), University of Cologne, Robert-Koch-Strasse 21, 50931 Cologne, Germany
2Cologne Excellence Cluster for Cellular Stress Responses in Aging-Associated Diseases (CECAD) Research Center and Systems Biology of Ageing Cologne, University of Cologne, Joseph-Stelzmann-Str. 26, 50931 Cologne, Germany
3Lead Contact
*Correspondence: mathieu.clement-ziza@uni-koeln.de (M.C.-Z.), aradaigl@uni-koeln.de (A.R.-I.)

http://dx.doi.org/10.1016/j.celrep.2016.11.046

SUMMARY

Cellular heterogeneity within embryonic and adult tissues is involved in multiple biological and pathological processes. Here, we present a simple epigenomic strategy that allows the functional dissection of cellular heterogeneity. By integrating H3K27me3 chromatin immunoprecipitation sequencing (ChIP-seq) and RNA sequencing (RNA-seq) data, we demonstrate that the presence of broad H3K27me3 domains at transcriptionally active genes reflects the heterogeneous expression of major cell identity regulators. Using dorsoventral patterning of the spinal neural tube as a model, the proposed approach successfully identifies the majority of previously known dorsoventral patterning transcription factors with high sensitivity and precision. Moreover, poorly characterized patterning regulators can be similarly predicted, as shown for ZNF488, which confers p1/p2 neural progenitor identity. Finally, we show that, as our strategy is based on universal chromatin features, it can be used to functionally dissect cellular heterogeneity within various organisms and tissues, thus illustrating its potential applicability to a broad range of biological and pathological contexts.

INTRODUCTION

During vertebrate embryogenesis, developing tissues display high cellular heterogeneity as they contain multiple progenitor and differentiating cellular states that are critical for tissue formation and function. Similarly, the homeostasis and regenerative potential of adult epithelial tissues depends on heterogeneous and distinct stem cell populations residing within such tissues (Donati and Watt, 2015). Therefore, understanding the molecular basis of cellular heterogeneity can provide major insights into tissue formation and homeostasis, which are typically altered in congenital and age-related diseases, respectively (Greene and Copp, 2014; López-Otín et al., 2013).

Some of the best-studied examples of functional cellular heterogeneity occur during embryonic patterning, a process whereby cellular progenitors within developing tissues acquire positional identity with respect to the major body axis (Wolpert, 1996). The regulatory principles of embryonic patterning in vertebrates are probably best understood along the dorsoventral axis of the spinal neural tube (SNT) (Cohen et al., 2013; Le Dréau and Martí, 2012), where the expression of specific combinations of transcription factors (TFs) (i.e., patterning TFs) define 11 distinct neural progenitor domains, each able to generate specific types of motors and interneurons. One distinctive feature of patterning regulators in general and patterning TFs in particular is their spatially confined expression within the tissues they regulate. Patterning regulators have been traditionally discovered using techniques, such as in situ hybridization and immunofluorescence, whereby the expression of pre-selected candidate genes is screened in tissues of interest, followed by functional characterization (Briscoe et al., 2000; Ericson et al., 1997). As a result, the identification of patterning regulators in vertebrates has been a slow, biased, and labor-intensive process, with most knowledge gathered for a few embryonic tissues (e.g., SNT, limb, and brain; Kiecker and Lumsden, 2005; Le Dréau and Martí, 2012; Zeller et al., 2009). In principle, these limitations could be overcome by using gene expression profiling from micro-dissected tissues or single cells (Jaitin et al., 2014; Junker et al., 2014). However, these approaches are technically challenging and are not designed to discriminate between genes that are simply expressed in a spatially confined manner and those that, additionally, have major regulatory functions within the investigated tissues.

Epigenomic profiling and the use of chromatin signatures have emerged as powerful and universal tools to functionally annotate genomes (ENCOD Project Consortium, 2012). Importantly, both genes and regulatory elements with major cell identity functions can be identified based on the presence of particularly broad chromatin features (Benayoun et al., 2014; Rada-Iglesias et al., 2011; Whyte et al., 2013). For example, master developmental genes are covered with broad H3K4me3 or H3K27me3
Figure 1. Epigenomic Strategy to Predict Dorsoventral Patterning Regulators in the Spinal Neural Tube

(A) Embryonic tissue subject to patterning in which certain genes are expressed in some cells (red cells) while being inactive and marked with H3K27me3 in others (blue and green cells). Genes expressed in such spatially restricted manner can be identified as transcriptionally active genes (as measured by RNA-seq) that are marked by H3K27me3. Among them, those covered by particularly broad H3K27me3 regions are likely to represent master developmental regulators controlling patterning of the investigated tissue.
domains when found in a transcriptionally active or inactive state, respectively (Benayoun et al., 2014; Ku et al., 2008; Nakamura et al., 2014). Here, we leverage the predictive power of chromatin signatures to develop a simple strategy to systematically identify major patterning regulators within embryonic tissues of interest. Patterning regulators are characterized by at least two major distinctive features: (1) they are expressed in a spatially restricted manner and (2) they can be broadly classified as core developmental regulators. We hypothesized that both features can be predicted by combining RNA sequencing (RNA-seq) and chromatin immunoprecipitation sequencing (ChiP-seq) profiling in embryonic tissues subject to patterning. Briefly, the presence of H3K27me3 at the promoter regions of transcriptionally active genes (as measured by RNA-seq) should reflect cellular heterogeneity and could be used to predict genes with spatially restricted expression (Akkers et al., 2009; Cotney et al., 2012; Riising et al., 2014; Schertel et al., 2015). Additionally, if the H3K27me3-marked regions are particularly broad, then the underlying genes should be enriched in master developmental regulators (Ku et al., 2008; Nakamura et al., 2014) that, given their spatially confined expression, could control embryonic patterning (Figure 1A).

RESULTS

Optimization of ChiP-Seq Protocols and Spinal Neural Tube Dissections in Chicken Embryos

The feasibility of our epigenomic strategy (Figure 1A) could be compromised by the technical difficulties in generating H3K27me3 ChiP-seq data from the limited biological material that can be isolated from embryonic tissues. Moreover, it was important to investigate whether our ChiP protocol could detect H3K27me3 enrichments for genes heterogeneously expressed within a cell population. To address these issues, tissue heterogeneity was first simulated in vitro by mixing two different cell lines (mouse embryonic stem cells [mESCs] and 3T3 mouse embryonic fibroblasts) in pre-determined ratios and overall low numbers (~6 × 10^5 cells). Using publically available RNA-seq and H3K27me3 ChiP-seq data (Bernstein et al., 2006), we selected a couple of genes displaying homogeneous but opposite H3K27me3 expression states in each of the two cell lines: SalI4 is active and negative for H3K27me3 in mESCs, whereas in 3T3 cells, it is inactive and covered by a broad H3K27me3 domain; Twist1 shows the opposite patterns (Figure S1A). H3K27me3 ChiPs were performed from samples containing a total of 6 × 10^5 mESCs and/or 3T3 cells mixed in different proportions. ChiP-qPCR assays showed that enrichments for H3K27me3 could be detected even when, at the investigated genes, this histone modification was only present in 20% of the mixed cell population (~1.2 × 10^5 cells; Figure S1B).

Next, potential sources of cellular heterogeneity within the SNT other than dorsoventral patterning had to be minimized, as they could affect the predictive power of our approach. These sources included (1) anteroposterior patterning, (2) coexistence of neural progenitors and differentiated neurons, and (3) contamination from surrounding non-neural tissues (Figure 1B). Heterogeneity due to anteroposterior patterning and neuronal differentiation was minimized by collecting SNT as narrow sections at a fixed rostrocaudal level (i.e., brachial level) from stage Hamburger and Hamilton (HH) 14 chicken embryos, which precedes the onset of neuronal differentiation (Ericson et al., 1992; Figure 1C). The contamination from surrounding tissues was minimized by briefly treating the dissected SNTs with trypsin, which facilitated the removal of non-neural tissues (Figure 1C). Overall, these measures limited the amount of biological material we could collect from each embryo. Thus, RNA-seq and H3K27me3 and H3K4me3 ChiP-seq experiments were performed using pooled SNT sections from HH14 chicken embryos as biological duplicates (~25–30 embryos [10^5–2 × 10^5 cells]/ChiP; Figures S1C–S1F; Data S1 and S2). As shown in Figure 1D, HOX gene clusters were sharply divided into active and inactive domains, suggesting limited anteroposterior heterogeneity within our SNT samples (Noordermeer et al., 2014; Figure 1D). Similarly, differentiated neurons were rare in our samples, as neuronal differentiation markers were marked by H3K27me3 and displayed low expression levels (e.g., ISL1; Figures 1E and 2D). Lastly, the dissected SNTs showed high expression of neural progenitor genes (e.g., SOX2) and virtually no expression of notochord (e.g., CHRD), paraxial mesoderm (e.g., TBX6), somites (e.g., MYOD1), or surface ectoderm (e.g., CXCL12) markers (Figures 1E and S1G).

Gene Expression Levels and H3K27me3 Breadth Can Be Combined to Functionally Dissect Cellular Heterogeneity

An initial evaluation of our RNA-seq and ChiP-seq datasets indicated that previously described SNT dorsoventral patterning TFs were frequently marked by broad H3K27me3 domains and displayed clear transcriptional activity (Figure 2A). Next, we systematically investigated a list of previously reported spinal tube dorsoventral TFs (n = 16) with respect to their expression levels and H3K27me3 breadth around their transcription start sites (TSSs) (Chizhikov and Millen, 2004; Cohen et al., 2013; Holz et al., 2010; Le Dréau...
Importantly, although these patterning TFs were transcriptionally active (e.g., 16/16 patterning TFs displayed transcripts per million [TPM] > 1), they were not expressed at particularly high levels (Figure 2B), suggesting that RNA-seq data alone cannot be used to functionally discriminate major patterning regulators. In contrast, most patterning TFs displayed distinctively broad H3K27me3 domains around their TSS (e.g., 15/16 patterning embedded within >5-kb-long H3K27me3 domains; Figure 2C), suggesting that the breadth of H3K27me3 could have superior predictive power to identify patterning regulators than RNA-seq. However, transcriptionally inactive genes involved in the development and patterning of non-neural tissues (e.g., TBX5), the anterior neural tube (e.g., OTX2), or in neuronal differentiation (e.g., ISL1) were similarly embedded within broad H3K27me3 domains (Figure 2D).

Based on these initial observations, we decided to combine the expression levels and H3K27me3 breadth of each gene into a single score, giving more weight to the latter. Briefly, a functional heterogeneity score (FH score) was calculated for each gene in the chicken genome by multiplying the breadth of the H3K27me3 domain (in bp) around its TSS by its expression levels (TPM in log scale). In principle, genes with high FH scores should represent major developmental regulators that are expressed by only a subset of progenitors within the dorsoventral axis of the SNT, thus leading to the simultaneous detection of H3K27me3 enrichment and transcriptional activity. To assess the predictive capacity of the FH score, we used a list of previously described SNT dorsoventral patterning TFs (Chizhikov and Millen, 2004; Cohen et al., 2013; Holz et al., 2010; Le Dreau and Marti, 2012; Li et al., 2014; Millonig et al., 2000; Thélue et al., 2015). Due to the incomplete annotation of the chicken genome, RNA-seq and ChIP-seq data were not available for some dorsoventral patterning TFs (e.g., OLIG2, NKX2-2, NKX6-1, and IRX3) that were therefore not taken into account. Finally, this list included 16 TFs (Data S2) that were considered as known dorsoventral patterning regulators with high specificity and sensitivity (A) ChIP-seq and RNA-seq profiles from HH14 chicken SNT around a representative dorsoventral patterning regulator (e.g., NKX6-2).

(B and C) All chicken genes were ranked according to (A) their expression levels (as TPM) or (B) H3K27me3 block sizes (in kb) in HH14 SNT. Previously known dorsoventral patterning TFs are shown in red. The red dashed lines indicate TPM = 1 (B) and H3K27me3 block size = 5 kb (C), respectively.

(D) ChIP-seq and RNA-seq profiles from HH14 chicken SNT are shown around a representative gene, driving neuronal differentiation (e.g., ISL1).

(E and F) Receiver operating characteristic (ROC) (E) and precision-recall (PR) (F) curves are shown for several alternative methods used to predict previously known SNT dorsoventral patterning TFs.

(G) Area under the curve (AU) values obtained for ROC (AUROC) and PR (AUPR) curves generated for several alternative methods used to predict previously known SNT dorsoventral patterning TFs.

(H) Distribution of TFs previously implicated in dorsoventral patterning of the SNT (total = 16) among the top 100 genes ranked according to FH score (blue), H3K4me3 breadth (orange), H3K27me3 breadth (purple), and gene expression (red). p values were calculated using hypergeometric tests.
with the highest FH scores (14/16; p = 7E–30; Figure 2 H).

ventral patterning TFs were found among the top 100 genes predictors of major patterning regulators, we ranked all
dorsoventral patterning regulators was identified (Figure 2 H).

were considered instead, a considerably smaller fraction of
to either H3K27me3 breadth, H3K4me3 breadth, or TPM
In stark contrast, when the top 100 genes ranked according
Remarkably, ~90% of the previously described SNT dorsoven-
tal patterning TFs were found among the top 100 genes with
the highest FH scores (14/16; p = 7E–30; Figure 2H).

In stark contrast, when the top 100 genes ranked according
to either H3K27me3 breadth, H3K4me3 breadth, or TPM
were considered instead, a considerably smaller fraction of
dorsoventral patterning regulators was identified (Figure 2H).

Among the regulators successfully identified by FH scores,
there were genes displaying either broad (e.g., PAX7) or narrow
(e.g., DBX1) expression domains in the SNT, thus illustrating
that our strategy can be used to identify genes within a
broad range of spatially restricted expression patterns.
Notably, the list of top 100 FH score genes was not limited
to patterning TFs but also included key morphogens involved
in the dorsoventral patterning of the SNT (e.g., SHH; Data S2).
It is worth noting that, due to slight technical variation during
SNT dissection, the top 100 FH score genes included a few
HOX genes.

Genes with High FH Scores Display Unique Functional
and Chromatin Features

Gene Ontology analyses of the top 100 FH score genes
showed a strong enrichment in developmental regulators
involved in dorsoventral patterning of the SNT (e.g., SHH, PAX6, DBX1, and NKX6-2; Figures 3A and 3B; Data S3).
In contrast, functional annotation of genes covered with
the broadest H3K27me3 domains revealed a significant
enrichment of patterning regulators (e.g., antero-posterior,
proximo-distal, and dorsoventral), regulators of non-neural
tissues (e.g., limb and muscle), and genes promoting neuronal
differentiation (Figures S2A–S2C; Data S3). Moreover, the
genes with the broadest H3K4me3 domains were not impli-
cated in SNT dorsoventral patterning and, instead, included
regulators of neural progenitor identity uniformly expressed
along the dorsoventral axis of the SNT (e.g., SOX2 and SOX3; Figures S2D–S2F; Data S3). Lastly, the genes with the
highest expression levels within the SNT were mostly enriched
in housekeeping functions, such as ribogenesis and energy
production (Figures S2G–S2I; Data S3).

The previous functional annotation analyses indicate clear
differences between high-FH-score genes and genes covered
by the broadest H3K27me3 domains. Accordingly, genes with
high FH scores were expressed within the SNT at significantly
higher levels than genes simply covered with broad H3K27me3
domains (Figure S3A). Similarly, the fraction of high-FH-score
genes enriched in H3K4me3 around their promoters (37%)
was considerably higher than among broad H3K27me3 genes
(5%; Figures 3A and S2A; Data S2). The failure to detect
H3K4me3 enrichments at some high-FH-score genes can be
the result of these genes being expressed in a small fraction
of cells and thus beyond the sensitivity of our H3K4me3
ChiP-seq or, alternatively, due to the absence of H3K4me3
at the promoter regions of some developmentally regulated
genes (Pérez-Lluch et al., 2015). Overlapping enrichments in
H3K27me3 and H3K4me3 have been previously described at
the promoters of genes broadly referred to as bivalent (Bern-
stein et al., 2006; Ku et al., 2008). These bivalent genes are
simultaneously occupied by these two histone marks within
the same cells and are transcriptionally inactive (Bernstein
et al., 2006; Schertel et al., 2015). In contrast, high-FH-score
genes should not be considered as bivalent, as the overlapping
enrichments in H3K27me3 and H3K4me3 at their promoters
arise due to cellular heterogeneity and spatially restricted
gene expression (Figure 2A). On the other hand, genes ex-
pressed in a cell-type-specific manner are frequently under
the control of distal regulatory sequences (i.e., enhancers;
Buecker and Wysocka, 2012). Putative enhancers active in
the SNT were identified using H3K27ac ChiP-seq, a histone
mark that characterizes this type of regulatory sequences
(Creyghton et al., 2010; Rada-Iglesias et al., 2011; Data S4),
as exemplified by the strong H3K27ac signals observed at
conserved enhancers previously linked to bona fide SNT dorso-
ventral patterning regulators (e.g., NKX6-2, DBX2, and NKX6-1;
Oosterveen et al., 2012; Figure S3B). Using a stringent win-
dow of 50 kb around TSS, we found that a significantly larger
fraction of high-FH-score genes (51%) were associated with
at least one distal H3K27ac peak compared to genes covered
by broad H3K27me3 domains (24%; Figure S3C).

Genes Predicted as Novel Spinal Neural Tube
Dorsoventral Patterning Regulators Display Spatially
Restricted Expression Patterns

Although the dorsoventral patterning of the SNT has been exten-
sively studied, we noticed that, among the top 100 FH score
genes, there were several genes that have not been previously
implicated in this patterning process. Therefore, we decided
to investigate whether our FH scoring system could be used
as a tool to uncover not only previously known but also novel
dorsoventral patterning regulators. To start evaluating this
possibility, we first tested the capacity of our FH scoring system
to detect heterogeneously expressed genes, as one major feature
of patterning regulators is their spatially restricted expression.
We selected ten candidates among the top 100 FH score genes
that had little or no evidence regarding a functional role during
patterning of the SNT (e.g., ZNF488, LMO1, PURA, and RELN) and investigated their expression patterns by in situ hybridization. Importantly, in nine of ten cases, we observed dorsoventrally restricted expression within the SNT (Figure 3C). Moreover, the selected high-FH-score genes were not biased toward particular dorsoventral domains, as they were expressed in dorsal (e.g., ZIC3), intermediate (e.g., ZNF488), or ventral (e.g., RELN) domains as well as in broad (e.g., LMO1) or narrow areas (e.g., NTN1).

**ZNF488 Expression Is Restricted to p1/p2 Interneuron Progenitor Domains**

Spatially restricted expression along the SNT dorsoventral axis is suggestive, but not proof, of patterning regulatory function. Hence, ZNF488 was selected as a candidate for further functional characterization (Figures 4A and S3C). The chicken ZNF488 protein contains two zinc-finger DNA-binding domains and a SET histone methyltransferase domain absent in its mammalian homologs (Wang et al., 2006), thus resembling members of the PRDM family (Hohenauer and Moore, 2012). Although some PRDM genes (e.g., Prdm12 and Prdm8) are known to be expressed within specific dorsoventral SNT progenitors (Komai et al., 2009; Thélie et al., 2015), the role of ZNF488 and most other PRDM family members during SNT patterning has been barely explored. Using RNA in situ hybridization, we found that, at stage HH14, ZNF488 was expressed together with OLIG2 in a mixed p2/MN progenitor domain (Chen et al., 2011; Figure 4B). At later stages, ZNF488 expression became more dorsal and clear expression boundaries emerged ventrally, with OLIG2, and dorsally, with DBX1 (Figures...
Figure 4. ZNF488 Is a Dorsoventral Patterning Regulator Candidate with Restricted Expression in p1 and p2 Ventral Progenitors

(A) ChIP-seq and RNA-seq profiles from SNT of HH14 chick embryos are shown around the ZNF488 gene.

(B and C) Dorsoventral expression of ZNF488 with respect to a panel of bona fide dorsoventral patterning TFs was investigated by in situ hybridization. Cross sections were all obtained at the brachial level of stage HH14 (B) or HH18 (C) chick embryos.

(D) Double in situ hybridization experiments were performed for ZNF488 (blue) and the indicated dorsoventral patterning TFs (red/orange). Cross sections were all obtained at the brachial level of stage HH18 chick embryos.
Figure 5. **ZNF488 Confers p1/p2 Ventral Progenitor Identity to Spinal Neural Tube Progenitors**

(A and B) Panel of bona fide dorsoventral patterning TFs whose expression was either affected (A) or unaffected (B) by ZNF488 overexpression in cross sections obtained at the brachial level of stage HH18 chicken embryos.

(legend continued on next page)
To directly test whether ZNF488 acts as a dorsoventral patterning regulator, we performed gain-of-function experiments whereby ZNF488 was overexpressed together with GFP into one-half of the SNT of stage HH12–13 chicken embryos (Figures S4A and S4B; Lacomme et al., 2012). Experiments in which GFP was overexpressed alone served as negative controls (Figure 5). The reduced GFP signals observed in some of the SNT sections shown in Figure 5 were the result of using digoxigenin NBT/BCIP complexes to detect the RNA in situ hybridization probes, which are known to mask fluorescence in those areas where the in situ signal is particularly strong (Lacomme et al., 2012). In agreement with its expression in p1/p2 interneuron progenitors, ZNF488 overexpression led to a strong repression of DBX1 and OLIG2 (Figures 5A and S4C). In contrast, NKX6-2 was expanded dorsally, most likely as a result of DBX1 repression, whereas DBX2 and NKX6-1 were not affected (Figures 5A, 5B, and S4C). In addition, we found that ZNF448 was able to repress NKX2-2 and NKX2-8, which are TFs characterizing the more ventrally located p3 interneuron progenitor domain (Figure S4D). Expression changes in dorsoventral patterning TFs can alter the differentiation of specific neuronal subtypes (Cohen et al., 2013; Le Dréau and Martí, 2012). Importantly, ZNF488 overexpression led to a major reduction in V0 interneurons, a dorsal expansion of V1 interneurons, and a ventral expansion of V2a interneurons, as reflected by the expression changes observed for EN1 and CHX10, respectively (Figures 5C and S4E). In addition, there was a reduction in motor neurons, as reflected by diminished expression of ISL1 and MNR2 (Figures 5D, 5E, S4F, and S4G).

To further test the specificity of our gain-of-function experiments and to gain some mechanistic insights into the function of ZNF488 as a patterning regulator, overexpression experiments were performed with truncated versions of ZNF488, in which the DNA-binding or SET domains were missing (Figure S5). In the absence of the SET domain and, especially, of the DNA-binding domain, the capacity of ZNF488 to repress their neighboring patterning TFs (e.g., DBX1 and OLIG2) was severely compromised (Figures S5A–S5C). These results support the role of ZNF488 as a SNT dorsoventral patterning regulator and suggest that, within the SNT, it preferentially acts as a TF.

Overall, our results suggest that ZNF488 represents a dorsoventral patterning regulator conferring p1/p2 interneuron progenitor identity (Figure 5F). Together with the successful prediction of ~90% of previously reported patterning TFs, this strongly supports the use of FH scores as a powerful tool to identify embryonic patterning regulators.

**Broad Applicability of FH Scores to Predict Major Developmental Organisms within Various Embryonic Contexts and Model Organisms**

Our epigenomic approach is based on simple and universal chromatin features, which in principle should make it applicable to other embryonic contexts and model organisms. To test its broad applicability, we first focused on the heterogeneity that emerges within the SNT as neural differentiation starts and that leads to the coexistence of undifferentiated progenitors and differentiating neurons (Le Dréau and Martí, 2012). We hypothesized that, at later developmental stages (e.g., HH19), genes representing major drivers of neuronal differentiation should become induced and, due to the presence of undifferentiated progenitors, should also be covered by broad H3K27me3 domains. To investigate this idea, we generated RNA-seq and H3K27me3 ChIP-seq data from stage HH19 SNTs, which represents an early stage of neuronal differentiation at the brachial level of the SNT (Data S1 and S5). In principle, major neuronal differentiation genes should be covered by broad H3K27me3 domains and transcriptionally active at HH19, while being silenced or expressed at lower levels at HH14. To detect these genes, we developed a modified version of the FH score, the differential FH score (dFH), which combined the H3K27me3 breadth with the gene expression fold-change levels between HH19 and HH14 (Figures 6A–6D; Data S5). Then, genes were ranked according to this dFH score, which revealed that as many as 17 out of the top 100 dFH score genes coincided with bona fide TFs promoting the differentiation of specific neuronal subtypes in the SNT (e.g., LHX1/3/4/5, ISL1, and PTF1A; Figures 6D and 6E; Data S5; Andersson et al., 2012; Blacklaws et al., 2015; Bröhl et al., 2008; Dubreuil et al., 2000; Glasgow et al., 2005; Jessell, 2000; Le Dréau and Martí, 2012; Roy et al., 2012; Zhou et al., 2000). Importantly, this enrichment in TFs promoting neuronal differentiation was considerably higher among the top 100 dFH scores genes than among genes with the largest expression differences between HH19 and HH14 SNT (Figure 6E), highlighting the predictive power of our dFH scoring approach. The presence of neuronal differentiation genes among high-dFH-score genes was not restricted to TFs, as there were additional genes with well-established neurotrophic roles and diverse molecular functions (e.g., NRN1, SLC32A1, CDK5R2, NTN5G2, and PLXND1; Pecho-Vrieseling et al., 2009; Sabatier et al., 2004; Zhou and Zhou, 2014).
Figure 6. Major Cell Identity Regulators within Heterogeneous Embryonic Tissues Can Be Efficiently Predicted Using the FH Scoring System

(A) Neuronal differentiation within the SNT starts after stage HH14 at the brachial level. By stage HH19, cellular heterogeneity has emerged due to the coexistence of undifferentiated neural progenitors and differentiating neurons.

(legend continued on next page)
To test whether our approach could be also useful in other embryonic tissues and model organisms, we took advantage of publicly available data. First, we used RNA-seq, H3K4me3 ChiP-seq, and H3K27me3 ChiP-seq data generated in embryonic day 10.5 (E10.5) mouse forebrains (Noordmeer et al., 2011, 2014; Figure 6F). At this developmental stage, cellular heterogeneity within the forebrain can arise due to (1) undifferentiated progenitors versus differentiated neurons, (2) anteroposterior patterning, and (3) dorsoventral patterning (Kiecker and Lumsden, 2005). Importantly, Gene Ontology analysis of the top 100 FH score genes indicated an overrepresentation of major forebrain regulators, including genes involved in forebrain patterning (e.g., *Otx2, Foxg1*, and *Emx2*), maintenance of undifferentiated neural progenitors (e.g., *Sox1, Sox2*, and *Pou3f2*), and neuronal differentiation (e.g., *Neurog2* and *Lhx2*; Figures 6G and 6A). Moreover, several genes with unknown function during forebrain development were found among this limited list, including various poorly characterized large intergenic noncoding RNAs (lincRNAs) and microRNA (miRNA) precursors (Data S6). Lastly, we analyzed RNA-seq and H3K27me3 ChiP-seq data generated in limbs isolated from E11.5 mouse embryos (Attanasio et al., 2014; Figure 6H). At this developmental stage, the limbs display considerable cellular heterogeneity due to (1) patterning (i.e., proximo-distal, anterior-posterior, and dorsoventral), (2) presence of undifferentiated mesenchymal progenitors, and (3) mesenchymal differentiation toward various cellular lineages (i.e., cartilage, bone, muscle, and vasculature; Zeller et al., 2009). Gene Ontology analysis of top 100 FH score genes revealed a significant overrepresentation of master regulators of limb morphogenesis, including genes involved in limb growth and patterning (e.g., *Tbx3, Tbx5, Fgf8, Hand2*, and *Hoxd13*) as well as mesenchymal differentiation toward the different cell lineages present in the limb, such as cartilage, muscle, or bone (Figures 6I and 6B; Data S6). The complexity and heterogeneity of the brain and the limbs at the investigated mouse developmental stages is considerably larger than within the HH14 chicken SNT. Consequently, the list of high-FH genes representing major cell identity regulators within these mouse embryonic tissues is likely to be longer than the top 100 FH score genes used here to illustrate the applicability and versatility of our approach. In agreement with this, when considering the top 500 FH genes in the limb, we observed a considerably larger number of genes known to be involved in limb development and mesenchymal differentiation, including master regulators of cartilage (e.g., Sox9), muscle (e.g., Myod1), or bone (e.g., Foxc1; Figure S6C).

**DISCUSSION**

Epigenomic approaches based on the use of chromatin signatures have greatly facilitated the functional annotation of vertebrate genomes (ENCODE Project Consortium, 2012). However, the profiling of histone modifications has been somehow restricted to in vitro cell lines and cell types that can be sorted in vivo in high quantities. There are at least two main reasons for this: (1) epigenomic profiling is based on ChiP-seq technology, which normally requires cellular material in amounts difficult to accumulate in vivo, and (2) histone modifications are pervasively used in all cell types. Hence, in adult and embryonic tissues with a heterogeneous cellular composition, it is not possible to discriminate from which cell type the histone modification signals are being generated. Here, we actually leverage this heterogeneity to establish a simple and, we believe, universal approach to functionally dissect cellular heterogeneity within embryonic and adult tissues. Using the dorsoventral patterning of the SNT in chicken embryos as a model, we demonstrate that FH scores can be used to identify embryonic patterning regulators with high sensitivity and specificity. The combination of just one ChiP-seq map and one RNA-seq transcriptome, together with a simple bioinformatic pipeline to calculate FH scores, was sufficient to predict ~90% of the TFs previously described as SNT dorsoventral patterning regulators (Cohen et al., 2013; Le Dréau and Marti, 2012). Importantly, despite the previous and extensive characterization of SNT dorsoventral patterning, our epigenomic approach also predicted poorly characterized regulators, as illustrated by ZNF488. We believe that, due to its technical simplicity, cost effectiveness, and the universal chromatin features in which it is based, our approach can be easily adaptable by other laboratories interested in other types of cellular heterogeneity. However, we would like to note that our approach is, by definition, biased toward the identification of cell identity regulators whose spatially restricted or heterogeneous expression is regulated at a transcriptional level. Therefore, major regulators whose heterogeneous expression and/or activity are post-transcriptionally regulated are likely to be missed by our approach (e.g., Gli TFs in dorsoventral patterning of the SNT).

The performance of our FH scoring method was evaluated in the context of SNT dorsoventral patterning by using ROC and AUC analysis (Figure S6D). The performance of our approach was evaluated using a variety of metrics, including sensitivity, specificity, and area under the curve (AUC) for ROC analysis. The performance of our approach was compared to other methods, such as gene set enrichment analysis (GSEA) and cluster analysis, and found to be superior in terms of sensitivity and specificity. The biological processes enriched by our approach were related to dorsoventral patterning, indicating that FH scores can be used to identify regulatory features that are not captured by other methods. Finally, we demonstrated the utility of our approach in the context of SNT dorsoventral patterning by using ROC and AUC analysis, which provided a comprehensive evaluation of the performance of our approach in this context.
PR analyses. The FH score achieved higher AUROC and AUPR than all the other evaluated predictors. Importantly, precision (and AUPR) is the most reliable performance measure when there is a highly skewed distribution of class (Ackermann et al., 2012; Saito and Rehmsmeier, 2015). This is definitely the case for SNT dorsoventral patterning, because the “gold standard” set comprised only 16 TFs among a total of ~16,000 genes. Moreover, as highlighted by our results, this list of 16 TFs is incomplete and additional SNT dorsoventral patterning regulators exist and remain to be discovered. Consequently, we presume that many currently unknown patterning regulators would be labeled as false positives in our PR analyses; thus, perfect precision (i.e., AUPR = 1) should not be expected in the SNT dorsoventral patterning model. Therefore, we consider that, given the previous limitations, the obtained AUPR results support the high precision of the FH scoring method to functionally dissect cellular heterogeneity.

Recent RNA-seq-based approaches in which transcriptional profiles can be obtained from single cells or minute tissue sections can be used to disentangle cellular heterogeneity (Jaitin et al., 2014; Junker et al., 2014). However, these approaches remain experimentally and bioinformatically challenging for most laboratories. Moreover, current single-cell RNA-seq approaches appear to have limited capacity to detect low-abundance transcripts as well as to accurately estimate gene expression levels (Boroviak et al., 2015). Therefore, single-cell RNA-seq might not be particularly suitable for the identification of certain cell identity regulators, such as TFs, which are frequently expressed at low or intermediate levels. More generally, gene expression levels alone do not necessarily provide information regarding the functional relevance of actively expressed genes. On the other hand, because our approach is directly applied to heterogeneous cell populations, it is likely to be less sensitive that single-cell RNA-seq when particularly rare populations are of major interest. Hence, we anticipate that our epigenomic strategy and the use of FH scores should be highly complementary to single-cell RNA-seq approaches, as their combination could simultaneously reveal major cell identity regulators and their precise spatial and/or temporal expression patterns within heterogeneous tissues.

The simple and universal chromatin features behind our epigenomic approach make it, in principle, broadly applicable not only to other embryonic patterning events but, more generally, to other biological contexts in which cellular heterogeneity is of functional relevance. For example, it is becoming evident that the homeostasis and regenerative capacity of most adult epithelial tissues relies on distinct adult stem cell compartments, which, despite their importance, remain poorly characterized at a molecular level (Donati and Watt, 2015). Combined with simple sorting strategies, we speculate that our approach should facilitate the identification of the major regulators that allow different adult stem cell compartments to establish and/or maintain their unique identities. Another major but perhaps more challenging application of our approach could be to molecularly dissect tumor heterogeneity (Meacham and Morrison, 2013). Future work from our and other laboratories will hopefully take advantage of the described epigenomic strategy, which, either alone or in combination with single-cell RNA-seq approaches, could aid in the functional and molecular dissection of biologically relevant cellular heterogeneity.

EXPERIMENTAL PROCEDURES

**Chicken Embryos**

Fertilized chicken eggs (white leghorn; Gallus gallus domesticus) were obtained from a local breeder (LSL Rhein-Main) and incubated at 37°C and 80% humidity in a normal poultry egg incubator (Typenreihe Thermo-de Lux). Following microsurgical procedures, the eggs were re-incubated until the embryos reached the desired developmental stages. The developmental progress was determined according to the staging system of HH (Hamburger and Hamilton, 1992).

**In Ovo Electroporation**

Electroporations were performed using stage HH12–13 chicken embryos. 3.5–4 mL of albumin were removed by using a medical syringe to lower the blastoderm and make the embryo accessible for manipulation. The eggs were windowed, and the extra embryonic membrane was partially removed in the region to be electroporated. The vectors expressing the genes of interest (pCLG-ZNF488-EGFP or pCLG-EGFP) were mixed with fast green solution (Sigma) at a 2:1 ratio to ease the detection of the injection site. Subsequently, the DNA-Fast Green solution was microinjected into the target site of the developing neural tube with the help of borosilicate glass capillaries and electroporated as previously reported (Cai et al., 2005; Scaal et al., 2004). The electroporations were placed on each side of the microinjected neural tube, and five square pulses of 80 V within 20 ms width were applied to each embryo using the Intracell TSS20 OVIDYNE Electroporator. Following electroporation, the eggs were sealed with tape and re-incubated until the desired developmental stages (HH18 and HH22).

**Dissection of Spinal Neural Tubes for RNA-Seq and ChIP-Seq**

Fertile chicken eggs were incubated at 37°C for 48 hr until they reached stage HH14 or HH19 (Hamburger and Hamilton, 1992). Embryos were isolated from the eggs, transferred into 20 mL of 1x PBS, and the extraembryonic membranes were removed. Transverse SNT sections were collected at the brachial level, extending caudally into the thoracic region (Figure 1C). Most of the tissues that are located laterally/ventrally to the neural tube (e.g., lateral plate mesoderm, endoderm, and aorta) were removed manually. Each section was immersed in warm (37°C) trypsin (Trypsin.EDTA solution 1:250 [X]; Gibco; Figure 1C). Trypsinization time was tightly controlled to avoid over-digestion and dispersion of neural tissue and to retain the structural integrity of the SNT. Once connective tissues became loosened near the outline of the neural tube, trypsin was inactivated by rinsing with cold (4°C) DMEM containing 10% fetal bovine serum (FBS) and samples were transferred into cold 1x PBS. Following trypsin treatment, mesenchymal and ectodermal tissues were removed manually using forceps. After isolation, SNT sections were pooled in a 1.5-mL tube, flash-frozen in liquid nitrogen, and stored at ~80°C. The purity of the isolated SNT sections was initially evaluated by qRT-PCR analysis using the primers described in Data S7.

**ChIP-Seq**

For each ChIP-seq experiment, ~25–30 SNT sections isolated from stage HH14 or HH19 chicken embryos were used. SNT sections were briefly homogenized in DMEM media with 10% FBS serum/1M sodium butyrate and cross-linked with 1% formaldehyde at room temperature with rotation for 15 min. Cross-linking was quenched with 0.125 M glycine, and neural tissue was rinsed with cold 1x PBS, re-suspended in 300 μL of sonication buffer (10 mM Tris, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-deoxycholate, and 0.5% N-lauroylsarcosine), and incubated at 4°C for 10 min. Chromatin was sonicated (Bioruptor plus; Diagenode) to generate DNA fragments of ~200–500 bp. Sonicated chromatin was incubated overnight with 5 μg of antibodies against H3K27me3 (39536; Active Motif), H3K27ac (Active Motif), or H3K4me3 (39159; Active Motif; Rada-Iglesias et al., 2011), followed by 4–6 hr incubation with Protein G Dynabeads (Life Technologies). Beads were washed four times with 1 mL of cold radioimmunoprecipitation assay buffer.
RNA-Seq
 Approximately 25 SNT sections collected at stages HH14 (as biological duplicates) and HH19 were used for each RNA-seq experiment, which we performed as previously described (Respue1a et al., 2016; see Supplemental Experimental Procedures for details).

FH and dFH Scores
 The FH score was calculated as the product of the log expression level and the H3K27me3 breadth. The FH score was computed using custom R scripts:

\[ FH_g = \ln(TPM_g) \times (H3K27me3B_3 + 1) \]

where TPM\(_g\) is the expression level of the gene \(g\) expressed as transcripts per million (see Supplemental Experimental Procedures) and H3K27me3B\(_3\) the breadth of the H2K27me3 mark at the gene \(g\). We added 1 to the H3K27me3 breadth to avoid mathematical issue with zeros. The dFH score was computed as the FH score but using the log2 gene expression fold change instead of the TPM:

\[ dFH_g = DE_g \times (H3K27me3B_3 + 1) \]

where \(DE_g\) is the log2 gene expression fold change of the gene \(g\) calculated via a differential gene expression analysis using DESeq2 (Love et al., 2014).

Statistical Analysis
 To evaluate the performance of the FH score in comparison to other predictors, we used the AUROC and the AUPR. The ROC curve shows how the proportion of correctly classified positive instances (true positive rate, also named sensitivity or recall) varies with the proportion of incorrectly classified negative instances (false positive rate [FPR; Obuchowski, 2005]). The AUROC represents (1) the average specificity across all sensitivities, (2) the average sensitivity across all specificities, and (3) the probability of a known dorsoventral patterning TF to have a higher score than a random gene (Hanley and McNeil, 1982). The precision-recall curve compares the fraction of retrieved positive instances to the fraction of true positives. The AUPR therefore reflects the average precision of the predictions across all recall thresholds (Davis and Goadrich, 2008).

ACCESSION NUMBERS
 The accession number for the ChIP-seq and RNA-seq data reported in this paper is GEO: GSE89606.

SUPPLEMENTAL INFORMATION
 Supplemental Information includes Supplemental Experimental Procedures, six figures, and seven data files and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2016.11.046.

AUTHOR CONTRIBUTIONS
 A.R.-I. conceived and designed the study; R.R. performed most of the experiments with support from S.C.-M., C.T., and E.M. M.C.-Z., M.N., P.F., and A.R.-I. performed the genomic analysis. S.C.-M. performed the statistical analysis. R.R., M.C.-Z., and A.R.-I. wrote the manuscript with input from the other authors.

ACKNOWLEDGMENTS
 We thank Elisa Marti and the A.R.-I. lab members for insightful comments and critical reading of the manuscript and Peter Numberg and Janine Altmüller for technical assistance with next-generation sequencing. Work in the A.R.-I. laboratory was supported by CMMC intramural funding, DFG research grants (RA 2547/1-1, RA 2547/2-1, and TE 1007/3-1), UoC Advanced Researcher Group grant, CECAD grant, and Fritz Thyssen Stiftung grant. M.C.-Z. work is supported by the German Ministry of Science and Education (BMBF) as part of the e:Med program (grant no. 01ZX1406) and by the CMMC career advancement program grant. E.M. was supported by CMMC intramural funding.

Supplemental Information

Received: June 2, 2016
 Revised: October 5, 2016
 Accepted: November 14, 2016
 Published: December 13, 2016

REFERENCES


