Science
AAAS

# Supplementary Materials for

## Synthetic regulatory reconstitution reveals principles of mammalian *Hox* cluster regulation

Sudarshan Pinglay *et al*.

Corresponding authors: Esteban O. Mazzoni, eom204@nyu.edu; Jef D. Boeke, jef.boeke@nyulangone.org

*Science* **377**, eabk2820 (2022)
DOI: 10.1126/science.abk2820

**The PDF file includes:**



**Other Supplementary Material for this manuscript includes the following:**

## Materials and Methods

### Yeast and *E. coli* strains and media

All yeast work was performed in strain BY4741 using standard media and growth conditions. Yeast transformations were performed using the Lithium acetate method as previously described (*52*). Plasmids were purified from bacterial cultures in Luria Broth (LB) supplemented with the appropriate antibiotic (12.5 µg/mL Chloramphenicol for source BACs, 25 µg/mL Kanamycin for all assemblons and 50 µg/mL Carbenicillin for all other plasmids). Plasmids purified from small scale bacterial cultures (5-10 mL) were used for all steps except for delivery to mESCs in which case DNA was purified from a large scale (500 mL) culture. Both bacteria and yeast were grown at 30°C.

A full list of yeast strains from this study is in Table S4. A full list of plasmids from this study is in Table S5.

### Yeast colony PCR

With the exception of the 134kb *SynHoxA* half-assemblon build (see relevant section in methods), all yeast colonies were genotyped by hand. Briefly, a single yeast colony was resuspended in 10-40 µl of 20 mM NaOH and placed in thermocycler. Yeast colony suspensions were boiled at 95°C for 5 min and then cooled to 4°C for at least 5 min before proceeding to PCR. 1ul of yeast lysate was used as template in a 10 µl GoTaq Green reaction (Promega M7123) with 0.25 µM of primers. PCR program: 95°C - 5min; 30X (95°C – 20s, 55°C – 90s, 72°C - 1min); 72°C – 5min; 4°C – hold. PCRs were separated on a 1% agarose gel containing ethidium bromide. 1kb Plus DNA Ladder (New England Biolabs N0550S) was used as a molecular weight standard.

### BAC recovery from yeast to *E. coli*

Plasmids were isolated from 5-10 mL yeast cultures either by alkaline lysis (*53*) or using Zymo Yeast Miniprep Kit I (Zymo Research D2001). Plasmids were then recovered into DH10B ElectroMax *E. coli* by electroporation (Invitrogen 18290015) following the manufacturer's instructions.

### Yeast CRISPR/Cas9 editing

Donors bearing the desired mutations were generated by overlap extension PCR using Q5 polymerase (New England Biolabs M0492S) Given the high efficiency of homologous recombination following the generation of a defined double strand break, markers are not required to be encoded in the donor. This allows for seamless, marker-free editing. All CRISPR modifications in yeast were made as previously described (*37, 54*). Target strains were first transformed with the Cas9 plasmid (pNA519 pRS413-TEF1p-Cas9) and subsequently transformed with gRNA plasmids and linear donor fragments generated by PCR. Correct clones were identified by colony PCR followed by sequencing. Yeast CRISPR guide sequences are

listed in Table S5. Donor sequences are listed in the supplementary tables associated with each assemblon.

Field Inversion Gel Electrophoresis (FIGE)

*SynHox* assemblon BACs were verified by digesting a ~250-500ng purified by alkaline lysis (*55*) from small scale (5-10 mL) saturated bacterial culture with PvuI-HF (New England Biolabs R3150S). Digestion reactions were carried out at 37°C for 3-24 hours. The entire reaction was separated using the CHEF-DR system (Biorad 1703670) on a 1% low melting temperature agarose gel (Lonza 50100) in 0.5X TBE. The system was programmed using the auto algorithm function for fragments between 2kb and 50kb. 500 ng of lambda monocut ladder (New England Biolabs N3019S) were used as a molecular weight standard. Gels were stained after separation in a 0.5 µg/mL solution of ethidium bromide in water for 20-30 min and destained in water for 20-30 min before imaging.

Design and build of 134kb *SynHoxA*

The 134kb *SynHoxA* assemblon was designed to cover the H3K27me3 domain at *HoxA* in mESCs (mm10 chr6:52151343-52285368). Corresponding rat coordinates (rn6 chr4:82120263-82339548) were identified using the UCSC genome browser Convert tool. The rn6 genome contains an erroneous duplication at *HoxA* between gaps in the assembly. We deleted this duplicated sequence *in silico* to arrive at the final 134kb *SynHoxA* sequence. Primers spanning the entire locus were designed using a custom script based on the Primer3 algorithm (*56*), yielding a list of 29628 unique 18-24bp primers that contain each of the 4 bases at a of 15% and meet cutoffs for primer dimer and hairpin formation scores. This list was then manually pruned to 28 primer pairs that cover the entire sequence with an average amplicon length of 4.5kb (range: 3.2kb-5.4kb) and an average overlap with the adjacent amplicon of 207bp (range: 363bp – 102bp). See Table S6 for coordinates of design.

BACs containing Rat *HoxA* (CH230-79B14 and CH230-454G2) were obtained from the BACPAC resources center. DNA was isolated by alkaline lysis from 5-10 mL of overnight culture in LB supplemented with 12.5 µg/mL chloramphenicol. PCR amplicons were generated using Kapa HotStart Readymix (Kapa Biosystems KK2602) supplemented with 1 M Betaine (Sigma B0300-5VL) using 10-25 ng of BAC DNA as template and 0.3 uM of primers in a 20 µl reaction. All amplicons were generated with an annealing temperature of 68°C except for #12 and #16, which were generated at 65°C. PCR program: 95°C - 5min; 30X (98°C – 20s, 65°C/68°C – 30s, 72°C - 3min); 72°C – 5min; 4°C – hold.

134kb *SynHoxA* was built by first constructing two half-assemblons (segments 1-14 and segments 15-28). Two µl of each PCR amplicon were used to generate a pool, which was then transformed into yeast strain BY4741 with appropriate linkers (generated by overlap extension PCR) to direct assembly into a linearized vector (100 ng of I-SceI-digested pLM453) as previously described. (*36*) Linkers contained unique restriction enzyme sequences (I-*Sce*I and *Asi*SI) to facilitate isolation of the insert sequence from the vector. The right linker for each of half assemblon also encoded a selectable *URA3* gene. In each assembly step, yeast transformants were first screened for the correct phenotypes (e.g. Ura+). Subsequently, colonies were tested for

the presence/absence of assembly junctions using spanning primers (sequences are listed in Table S7) across the construct with the aid of a robotic workcell as previously described (*36*). Plasmids from correct yeast clones (ySP0084 and ySP0085) were recovered into *E. coli* by electroporation to generate larger quantities of DNA.

To build the full 134kb *SynHoxA* assemblon, half-assemblon BACs (pSP0180 and pSP0182) were purified by alkaline lysis. Inserts were released using AsiSI digest and ~1 μg of each were transformed into BY4741 with appropriate linker fragments and linearized vector (I-SceI digested pLM453). The right linker for this step encoded a *LEU2* marker, which allowed for the simple exclusion of yeast colonies arising from the transformation of undigested half-assemblon BACs (marked with *URA3*). Leu+ yeast colonies were screened for the presence of a PCR amplicon spanning segments 14-15 junction and for primers spanning the entire construct. Plasmids from correct yeast clones (ySP0088/89) were recovered to *E. coli* (pSP0193/196) and purified by alkaline lysis for verification by FIGE and sequencing.

In order to functionalize the assemblon for delivery by ICE, the backbone was modified in yeast using CRISPR. Strain bearing the full 134kb assemblon was transformed with Cas9 expressing plasmid pNA519 to yield yeast strain ySP0093. ySP0093 was subsequently transformed with a gRNA encoded in pSP0197 and a donor amplified from pLM707 containing the PGK-ATG-loxM-loxP-GFP cassette. The insertion was verified by Sanger sequencing. Plasmid was recovered from this yeast strain (ySP0096) into bacteria (pSP0211) and used for delivery to mESCs.

Table S7 lists all sequences used in the build of 134kb *SynHoxA*, including segment primers, junction primers, linkers, gRNAs and reagents used for amplification of the donor used to insert the ICE cassette.

Design and build of 170kb *Enhancers+SynHoxA*

Mouse enhancer coordinates were derived from published reports of enhancer function (*20*-*22*) and were expanded by adding 1 kb on each side. Mouse coordinates were then mapped to the rat genome using the UCSC genome browser Convert tool. In order to be conservative, we defined the final rat enhancer coordinates to be the union of those derived from the original and extended mouse enhancer coordinates (See Table S8 for details). All enhancer sequences were then appended to yield the compound enhancer sequence. Primers were designed to amplify the requisite sequences as well as to generate linkers between non-contiguous segments (Table S9).

PCR amplicons corresponding to the enhancer segments were generated using Q5 polymerase (New England Biolabs M0492S), except for segment 2 which was generated with Kapa polymerase supplemented with 1 M Betaine. 10-25 ng of CH230-79B14 BAC prepped by alkaline lysis was used as template with 0.25 μM primers in 20 μl reactions. All amplicons were generated at an extension temperature of 68°C except for segment 2, which was generated at 72°C. PCR program: 95°C - 5min; 30X (98°C – 20s, 65°C – 30s, 68/72°C - 3min); 72°C – 5min; 4°C – hold.

PCR amplicons corresponding to the linkers between non-contiguous enhancer segments were generated by overlap extension PCR. Initial amplicons were generated with Q5 polymerase supplemented with 1 M Betaine. These were gel purified using Zymoclean Gel DNA Recovery Kit (Zymo Research D4002) and the two fragments were mixed together at a 1:1 ratio and 1-10 ng of this mix was used as template with outer primers in 20 µl Q5 reactions to generate the final linkers.

10 µl of each segment's PCR product and linker product were pooled separately and then isopropanol-precipitated to a final volume of 10 µl of TE.

Yeast strain bearing the 134kb *SynHoxA* assemblon with delivery cassette (ySP0096) was transformed with Cas9 plasmid pNA519 to yield ySP0108. ySP0108 was transformed with plasmid pSP0233, which expresses a gRNA that cuts upstream of the 134kb assemblon, and with 10 µl of segment PCR pool and 5 µl of linker PCR pool to repair the gap. Yeast colonies were screened manually with junction primers spanning the overlaps between segments and with primers that span the 134kb *SynHoxA* assemblon (see Table S9 for enhancer junction primers and Table S14 for primers that span the 134kb *SynHoxA* assemblon).

A verified yeast colony (ySP0109) was recovered to bacteria (pSP0242) for verification by FIGE and sequencing.

Design and build of 130kb *RARE*Δ *SynHoxA* and 166kb *Enhancers+RARE*Δ *SynHoxA*

Rat RARE sequences were based on published reports of mouse RAREs that were converted to Rat coordinates using the UCSC genome browser Convert tool (*9*). The region containing 3'Hoxa1 RARE (rn6 chr4:82120263-82123922) was deleted. Other RARE mutations were designed such that only the direct repeat sequences were mutated to polyA. Mutagenesis was performed using CRISPR/Cas9. Donors were generated by overlap extension PCR (Table S10).

Hoxa7 CDS mutation R131W was corrected in ySP0108 (yeast strain with 134kb *SynHoxA* – ICE delivery cassette + pNA519 Cas9 plasmid) by CRISPR/Cas9. ySP0108 was transformed with gRNA expression plasmid pSP0255 and a donor generated from a synthetic oligo obtained from IDT (Table S10). Yeast colonies were screened by PCR followed by Sanger sequencing to yield wild type Hoxa7 strain ySP0126, which was transformed with Cas9 expression plasmid pNA519 to yield ySP0128, which is the parent to all subsequent assemblons.

3'Hoxa1 RARE was first deleted using gRNA expression plasmid pSP0277 to yield ySP0131. ySP0131 was transformed with pNA519 Cas9 expression plasmid to yield ySP0146. 5'Hoxa3 RARE and 5'Hoxa4 RAREs were modified in ySP0146 with gRNAs expressed from pSP0322 and requisite donors to yield ySP0147. Finally, 3'Hoxa4 RARE was mutated in ySP0147 using the gRNA expression plasmid pSP0315 to yield ySP0161. All mutations were verified in yeast colonies using genotyping primers specific to the mutation (Table S10) followed by Sanger sequencing of the region,

To add enhancer sequences to the 130kb *RARE*Δ *SynHoxA* assemblon, enhancers were amplified from the 170kb *Enhancers+SynHoxA* BAC (see section on build of 170kb assemblon). A pool of

enhancer amplicons was transformed into ySP0161 with gRNA expression plasmid pSP0233 and linkers to repair the cut upstream of the assemblon. Yeast colonies were screened for the presence of all enhancer junctions as well as sequences spanning the 134kb S*ynHoxA* assembly (see Table S9 for enhancer junction primers and Table S14 for primers that span the 134kb *SynHoxA* assemblon) The plasmids from ySP0161 and ySP0200 were recovered to bacteria and subject to verification by FIGE and sequencing to yield pSP0328 and pSP0394, respectively.

mESC culture, media and differentiation

A17iCre mESCs (*38*) were cultured on plastic tissue culture plates coated with 0.1% gelatin (EMD Millipore ES-006-B) at 37°C and 5% CO2 in '80/20' medium comprising 80% 2i medium and 20% mESC medium. mESCs were grown at 37°C and 8% CO2 before differentiation experiments.

2i medium was made from a 1:1 mix of Advanced DMEM/F12 (Gibco 12634010) and Neurobasal-A (Gibco 10888022), containing 1X N-2 supplement (Gibco 17502048), 1X B-27 supplement (Gibco 17504044), 2 mM Glutamax (Gibco 35050061), 0.1 mM Beta-Mercaptoethanol (Gibco 31350010), $10^3$ units/mL LIF (Millipore, ESG1107), 1 μM MEK1/2 inhibitor (Stemgent, PD0325901), and 3 μM GSK3 inhibitor (R&D Systems, CHIR99021). mESC medium was made from Knockout DMEM (Gibco 10829018), containing 15% Fetal Bovine Serum (Gemini), 0.1 mM Beta-Mercaptoethanol (Gibco 31350010), 1X MEM Non Essential Amino Acids (Gibco 11140050), Glutamax (Gibco 35050061), 1X Nucleosides (Millipore ES-008-D) and LIF (Millipore ESG1107).

The protocol for mESC *in vitro* differentiation to motor neurons has been described previously (*42*). Briefly, trypsinized (Gibco) mESCs were plated in AK medium (Advanced DMEM/F12:Neurobasal (1:1) medium (Gibco), 7% KnockOut SR (vol/vol) (Gibco), 2 mM L-glutamine, 26.5 μM ß-mercaptoethanol and penicillin–streptomycin (Gibco)) to induce formation of embryoid bodies (EBs), at 37°C, 5% CO2. $3.5x10^5$ cells were plated in 100 mm suspension plates (Corning) for RNA-seq experiments, while $3.5x10^6$ cells were plated in 245 mm x 245 mm square plates (Corning) for ChIP-seq experiments. After 2 days (0h timepoint), the EBs were split 1:2 and plated in fresh AK medium supplemented with 1 μM all-trans retinoic acid (RA) and 0.5 μM smoothened agonist (SAG) (Millipore 566660). For CHIR differentiation, in place of RA/SAG, CHIR99021 was added at 3 μM concentration

Table S11 lists all cell lines used in this study.

PCR genotyping of mESCs

mESC clones were initially screened by performing PCR on crude gDNA extracted from cells growing in a 96-well plate. Briefly, cells were washed once with PBS and frozen at -80°C for at least 30 min. The plate was thawed at room temperature and cells were resuspended in 40-50 μl of TE supplemented with 0.3 μg/ul Proteinase K (Thermo Scientific EO0492) and transferred to a 96-well PCR plate. Cell suspensions were then incubated in a thermocycler at 37°C for 1 hour and then 99°C for 10 min. 1 μl of this crude lysate was used for PCR in a 10 μl GoTaq Green reaction (Promega M7123) with 0.25 μM of primers. Candidate clones were expanded and re-

genotyped using DNA extracted with the QIAamp DNA mini kit (Qiagen 51306) according to the manufacturer's protocol. 25-50 ng of DNA was used as template per reaction. PCRs were separated on a 1% agarose gel containing ethidium bromide. 1kb Plus DNA Ladder (New England Biolabs N0550S) was used as a molecular weight standard.

Generation of *HoxA*$^{-/-}$ cell line

Deletion of the endogenous *HoxA* alleles was designed to mimic the sequence of 134kb *SynHoxA* (mm10 chr6:52151380-52285416). gRNAs were designed using CRISPOR and were chosen to have high cutting efficiency and low off-target scores (*57*). gRNAs were cloned into pSP0172, a modified version of pX459 (Addgene #62988, a gift from Feng Zhang) that has the Puromycin resistance gene replaced by a Blasticidin resistance gene. A 200 bp single stranded oligo donor (ssODN, oSP379) was designed to bridge the deletion and ordered from Integrated DNA Technologies (IDT).

gRNA expression plasmids were purified using PureLink HiPure Plasmid Midiprep Kit (Invitrogen K210004) following manufacturer's instructions. 12.5 µg of gRNA expression plasmids pSP0161 and pSP0164 were nucleofected into 1 million A17iCre mESCs along with 5 µl of 100 µM ssODN. Cells were transiently selected with 10µg/mL Blasticidin (Invivogen ant-bl-05b) for 3 days post nucleofection. Single clones were picked and genotyped using primers that span the deletion junctions and with primers internal to mouse *HoxA*. Sanger sequencing confirmed the precise deletion. Clones with the correct genotype were expanded and subject to WGS. Clones were also verified by metaphase spread karyotyping and quantitative real-time PCR (qRT-PCR) for ESC markers as previously described (*36*). A passing clone (1-A6) was used for all experiments.

Sequences of guides, donor and genotyping primers can be found in Table S12. Sequences of qRT-PCR primers are in Table S13.

Delivery of assemblons to mESCs and verification

Bacterial strain carrying the desired assemblon was struck out and grown for 2 days at 30°C on LB-agar plates containing 50 µg/mL Kanamycin. Single colonies were picked into 5 mL of LB + 25 µg/mL Kanamycin and grown for approximately 8 hours. The starter culture was used to seed 500 mL cultures at a 1:1000 dilution. 500 mL cultures were shaken at 30°C for 18-24 hours. Cells were pelleted and left at -20°C for long term storage or directly used for isolating DNA. BAC DNA was isolated using Nucleobond XtraBAC kit (Takara 740436.25) according to manufacturer's protocol. DNA pellets from 500 mL cultures were resuspended in 40 µl of TE.

DNA was delivered by nucleofection using the Nucleofector 2b system (Lonza VPH-1001). 1.5 million A17iCre mESCs were plated on gelatin-coated 10 cm dishes 2 days before delivery. Cre expression was induced with 1 µg/mL doxycycline 18-24 hours before delivery. On the day of delivery, cells were washed with PBS and harvested using Accutase (Biolegene 423201). 5-6 million cells were used per delivery. 10 µl purified assemblon DNA were used per delivery. Cells were resuspended in 100 µl nucleofection solution with purified DNA and transferred to a cuvette using wide bore tips. Cells were nucleofected using program A-023 and plated on

gelatinized 10 cm dishes. Cells were selected with 400 µg/mL Geneticin (Life Technologies, 10131-027) 48 hours post nucleofection. Resistant clones were picked 10-14 days post nucleofection. Cells were not kept under Geneticin selection after this point.

Crude gDNA was extracted from clones and PCR genotyping was performed with primers spanning genome junctions (tetO-GFP and PGK-Neo), primers specific to endogenous *HoxA* deletion and primers specific to the overwritten *Cre* in the landing pad. In addition, clones were screened with a subset of heterologous primers that were designed using Primer-Blast to be specific to *SynHoxA* sequences when compared to endogenous mouse *HoxA*. Correct clones were expanded, pure genomic DNA was extracted and clones were genotyped with the full complement of primers. Passing clones were then verified to contain the full assemblon by capture sequencing. Sequences of genotyping primers are in Table S14.

To generate the flow cytometry plot in Fig. S5, parental WT mESCs and capture-seq verified WT mESCs bearing the 134kb *SynHoxA* assemblon at *Hprt* were treated with 3 µg/mL doxycycline for 24 hours. Flow cytometry was performed on a BD Accuri C6 instrument and results were analyzed using FlowJo. Cells were gated on forward and side-scatter and histograms of GFP expression normalized to mode were plotted.

Library preparation for next generation sequencing

A list of all sequencing libraries and information associated with them can be found in Table S15.

*BAC DNA sequencing*

Illumina sequencing libraries were generated from 100-200 ng BAC DNA using NEBNext Ultra II FS DNA Library Prep Kit (New England Biolabs E7805S) according to the manufacturer's protocol. Libraries were sequenced on an Illumina NextSeq 500 in paired end mode.

*Sequencing mESCs (WGS and Capture-Seq)*

mESC gDNA was purified from 1-5 million cells using the QIAamp DNA mini kit (Qiagen 51306) according to the manufacturer's protocol. Illumina libraries were prepared as previously described (*41*). 1 µg of DNA was sheared to ~500 to 900 bp in a 96-well microplate using the Covaris LE220 (450 W, 10% Duty Factor, 200 cycles per burst, and 90-s treatment time). Sheared DNA was purified using the DNA Clean and Concentrate-5 Kit (Zymo Research D4013), and the concentration was measured on a NanoDrop instrument (Invitrogen). DNA fragments were end-repaired with T4 DNA polymerase, Klenow DNA polymerase, and T4 polynucleotide kinase (New England Biolabs M0203S, M0210S and M0201S, respectively), and A-tailed using Klenow (3′-5′ exo-; New England Biolabs M0212L). Illumina-compatible adapters were subsequently ligated to DNA ends, and DNA libraries were amplified with KAPA 2X Hi-Fi Hotstart Readymix (Roche).

Libraries for whole genome sequencing (WGS) of parental mESC lines (WT and *HoxA*$^{-/-}$) were sequenced on a Novaseq 6000 in paired end mode.

Targeted sequencing using in-solution hybridization capture (Capture-seq) was performed on *SynHoxA*-delivered mESCs as previously described (*41*). Biotinylated baits for capture sequencing were prepared from assemblon BACs using nick translation. 134kb *SynHoxA* mESCs were captured with bait made from 134kb *SynHoxA* BAC. All other mESCs were captured with bait made from 170kb *Enhancers+SynHoxA* BAC. In addition, the parental mESCs and RAREΔ *SynHoxA* mESCs were captured with a bait that included the ICE landing pad and flanking mouse genome sequence (pLM1103+ICEFlanking). See Table S15 for details. All libraries from *SynHoxA* deliveries were sequenced on an Illumina NextSeq 500 in paired end mode.

*ChIP-seq*

Cells were collected at 0h or 24h, 48h, or 96h after RA/SAG treatment. ChIP-seq was performed as previously described (*45*).

Cells were crosslinked at room temperature in 1 mM DSG (ProteoChem) for 15 min, followed by the addition of 1% FA (vol/vol) for 15 min. The reaction was quenched with Glycine and cells were washed with 1 × PBS. Samples were divided into ~25-30 million cell aliquots, pelleted by centrifugation at 275 g, and frozen at -80°C. Cell aliquots were thawed on ice and lysis was performed in 5 mL of 50 mM HEPES-KOH pH 7.5, 140 mM NaCl, 1 mM EDTA pH 8.0, 10% glycerol (vol/vol), 0.5% Igepal (vol/vol), 0.25% Triton X-100 (vol/vol) with 1 × protease inhibitors (Roche, 11697498001) for 10 min at 4°C. Cells were pelleted by centrifugation for 5 min at 1200 g, resuspended in 5 mL of 10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0 with 1 × protease inhibitors, and incubated for 10 min at 4°C on a rotating platform. Cells were centrifuged for 5 min at 1200 g and resuspended in 2 mL of Sonication Buffer (50 mM Hepes pH 7.5, 140 mM NaCl, 1 mM EDTA pH 8.0, 1 mM EGTA pH 8.0, 1% Triton X-100 (vol/vol), 0.1% sodium deoxycholate (wt/vol), 0.1% SDS (vol/vol) with 1 × protease inhibitors). For sonication, each sample was split in two Bioruptor tubes with added sonication beads. Sonication was performed using the Bioruptor Pico (Diagenode) for 18 cycles of 30 sec on and 30 sec off to sheer DNA into an average size of approximately 200bp. Immunoprecipitation was performed for 16h at 4°C on a rotating platform by incubating with Dynabeads protein-G (Thermo Fisher Scientific) conjugated with antibodies. For histone modifications, half of each original cell aliquot was incubated with Dynabeads protein-G conjugated with 5 µg of rabbit polyclonal to Histone H3K27me3 (Active motif 39155) or rabbit polyclonal to Histone H3 (acetyl K27) (Abcam ab4729) antibodies. For CTCF, the entire cell aliquot was incubated with Dynabeads protein-G conjugated with 5 µl of rabbit polyclonal to CTCF (EMD 07-729) antibody.

After the immunoprecipitation, washes were performed with the following ice-cold buffers: sonication buffer, sonication buffer with 500 mM NaCl, LiCl wash buffer (20 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 250 mM LiCl, 0.5% Igepal (vol/vol), 0.5% sodium deoxycholate (wt/vol)), and TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8). Elution was performed in elution buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA pH 8.0, 1% SDS (vol/vol)) by incubating for 45 min at 65°C with occasional flicking of the tube. Samples were incubating for 16h at 65°C to perform reversal of crosslinks. 200 µL of TE and RNase A (Sigma) at a final concentration of 0.2 mg/mL was added to digest RNA by and incubating for 2h at 37°C. Proteins were digested by adding Proteinase K (Invitrogen) at a final concentration of 0.2 mg/mL, supplemented with

CaCl2, at 55°C for 30 min. DNA was purified with phenol:chloroform:isoamyl alcohol (25:24:1; vol/vol) (Invitrogen) followed by an ethanol precipitation. DNA pellets were resuspended in 70 µL of water. lllumina DNA sequencing libraries were prepared with approximately one third of the ChIP sample (24 µL) or a 1:100 dilution of the input sample in water. Library preparation was performed by end repair, A-tailing and ligating Illumina-compatible Bioo Scientific multiplexed adapters. Agencourt AmpureXP beads (Beckman Coulter) were used to remove unligated adapters. PCR amplification was performed with Phusion polymerase (New England Biolabs) and TruSeq primers (Sigma). Libraries were gel purified (Qiagen) between 250 and 550bp in size. Libraries were quantified before pooling using the KAPA library amplification kit on the Roche Lightcycler 480 or the Bio-Rad CFX96. The libraries were sequenced on Illumina NextSeq 500 using V2.5 chemistry (75 cycles, single-end 75bp) or on Illumina NovaSeq 6000 using the SP Reagent Kit (100 cycles, single-end 100bp) at the Genomics Core Facility at NYU.

*RNA-seq*

Cells were collected at 0h or 24h, 48h, or 96h after RA/SAG treatment. RNA-seq was performed as previously described(*45*). RNA was extracted with the TRIzol LS Reagent (Life Technologies) followed by purification with the RNAeasy mini kit (Qiagen). RNA integrity was checked with the Agilent High Sensitivity RNA Screentape (Agilent, 5067-5579). 500 ng of RNA was used to prepare libraries and spiked-in with ERCC Exfold Spike-in mixes (Thermo Fisher 4456739). The TruSeq Stranded mRNA Library Preparation kit (Illumina 20020594) was used to prepare RNA-seq libraries. Library size was checked on the High Sensitivity DNA ScreenTape (Agilent 5067-5584). The KAPA library amplification kit was used to quantify libraries on the Bio-Rad CFX96 or the Roche Lightcycler 480 before pooling libraries. Libraries were sequenced on the Illumina NextSeq 500 using V2.5 chemistry (75 cycles, single-end 75 bp) or on the Illumina NovaSeq 6000 using the SP Reagent Kit (100 cycles, single-end 100bp) at the Genomics Core Facility at NYU. Control (without synthetic Hox constructs) RNA-seq datasets were previously published: 0h in (*44*); and 48h/96h after RA/SAG in (*45*).

*Hi-C*

Cells were collected at 0h and 48h after RA treatment. Cells were divided into ~$1x10^6$ aliquots and crosslinked in a final concentration of 2% FA (vol/vol) in 1 × PBS for 10 min at room temperature. After quenching with Glycine, cells incubated on ice for 15 min, pelleted by centrifugation at 500 g, and frozen at -80°C. Hi-C was performed using the Arima-HiC workflow (Arima Genomics, San Diego, CA) by NYU Langone's Genome Technology Center (RRID: SCR_017929).

Sequencing data analysis

*Custom references*

Two modified versions of the mm10 genome and corresponding genome annotations, were created using the reform tool (https://reform.bio.nyu.edu/). The genome mm10_synHoxA was created by replacing mm10 chrX:52963048-52997452 (Hprt) with the sequence corresponding to 170kb *Enhancers+SynHoxA* delivered to the ICE landing pad. A second genome

mm10_synHoxA_delHoxA was then created by removing the endogenous *HoxA* sequence that is deleted in the *HoxA*$^{-/-}$ mESC line (mm10 chr6:52151380-52285416).

The custom genome sequences, annotations and bowtie2 references can be found at: https://genome.med.nyu.edu/public/boekelab/SynHox_genomes/

*Parental mESCs WGS analysis*

WGS data were analyzed as previously described (*41*). Reads were demultiplexed with Illumina bcl2fastq v2.20 requiring a perfect match to indexing BC sequences. Illumina sequencing adapters were trimmed with Trimmomatic v0.39 (*58*). Reads were aligned to mm10 using BWA v0.7.17 (*59*). PCR duplicates were marked using samblaster v0.1.24 (*60*). Generation of per base coverage depth tracks and quantification was performed using BEDOPS v2.4.35 (*61*). Data were visualized using the University of California, Santa Cruz Genome Browser at coordinates mm10 chr6:52001889-52321810.

*Assemblon BAC sequencing analysis*

Reads were demultiplexed with Illumina bcl2fastq v2.20 requiring a perfect match to indexing BC sequences. Illumina sequencing adapters were trimmed with Trimmomatic v0.39. Reads were then mapped to mm10_synHoxA_delHoxA using bowtie2 v2.2.9 unless specified otherwise (*62*). Samtools was used to sort bam files and coverage tracks were generated using deeptools bamCoverage v3.2.1 with bin size set to 1, ignoring duplicates (*63*, *64*). Coverage was visualized in the Integrated Genomics Viewer (IGV) at coordinates chrX:52967270-53136430 (*65*).

Variants in *SynHoxA* assemblons were called relative to the rn6 rat reference genome. Sequencing data from assemblon BACs were mapped using the same pipeline described for parental mESCs (*41*). A modified rn6 genome was used as reference. Two sequences were masked: 1) the sequence corresponding to the mistaken duplication at *HoxA* (rn6 chr4: 82229539-82315425) and 2) an unplaced contig (4_KL567939v1_random) containing *HoxA* sequence. Variants were then called using a standard pipeline based on bcftools v1.9:

bcftools mpileup–redo-BAQ–adjust-MQ 50–gap-frac 0.05–max-depth 10000–max-idepth 200000 -a DP,AD–output-type u |

bcftools call–keep-alts –ploidy 1–multiallelic-caller -f GQ–output-type u
Raw pileups were filtered using:

bcftools norm–check-ref w–output-type u |

bcftools filter -i “INFO/DP>=10 & QUAL>=10 & GQ>=99 & FORMAT/DP>=10”–SnpGap 3–IndelGap 10–set-GTs .–output-type u |

bcftools view -i 'GT=”alt”"–trim-alt-alleles–output-type z

*Capture-sequencing coverage analysis*

All capture-seq data coming from assemblon delivery to WT mESCs was mapped to mm10_synHoxA whereas all data from *HoxA*$^{-/-}$ mESCs was mapped to mm10_synHoxA_delHoxA for coverage analysis. Reads were demultiplexed with Illumina bcl2fastq v2.20 requiring a perfect match to indexing BC sequences. Illumina sequencing adapters were trimmed with Trimmomatic v0.39. Reads were mapped using bowtie2 v2.2.9. Samtools was used to sort bam files and coverage tracks were generated using deeptools (*63*) bamCoverage v3.2.1 with bin size set to 1, ignoring duplicates. Coverage was visualized in the Integrated Genomics Viewer (IGV) coordinates chrX:52967270-53136430 (*65*).

*Capture-sequencing integration site analysis*

Bamintersect analysis was performed as previously described (*41*) with a few modifications. Briefly, capture-seq data were mapped to mm10 and independently to references containing the ICE landing pad sequence and the delivered assemblon sequence. Bamintersect identifies junctions by looking for read pairs where each read is mapped to a different reference.

To exclude spurious hits, certain sequences present in multiple contexts were masked. For example, the mouse Pgk1 promoter is found in the rtTA cassette integrated at Rosa26, at its endogenous location on chrX and at the 3' end of the delivered assemblon driving G418 (Neo) resistance. Similarly, the SV40 pA signal is present downstream of both the GFP in the assemblon and G418$^{R}$ (Neo) gene in the landing pad. Coordinates of masked sequences are: 1) Pgk1 promoter found in the rtTA cassette integrated at Rosa26 (mm10 chr6:113071694-113077114) 2) Pgk1 promoter – endogenous location (mm10 chrX:106186732-106187231) and 3) region of the mouse genome immediately downstream of the G418$^{R}$ SV40 pA (chrX:52962425-52963047). In addition, the endogenous mouse *HoxA* sequence (mm10 chr6:52121869-52285368) was masked to eliminate hits that arise from cross mapping of highly conserved regions between the rat derived assemblon and mouse *HoxA*.

Reads with the same strand and mapping to within 500 bp of each other were clustered for reporting. Regions below 150 bp or with fewer than 2 reads/10M reads sequenced were excluded.

*ChIP-seq data analysis*

Fastq files were aligned to the mm10_synHoxA_delHoxA custom genome using Bowtie 2 (*62*), using options -p 20. Samtools (*64*) was used to create sorted bam files for inputs into the bamCoverage tool from deeptools (*63*) to create bigWig files, using options: --binSize 1 --scaleFactor 0.001 --normalizeUsingRPKM --ignoreForNormalization chrX -p max --extendReads 100. The bigWig files were visualized using IGV (*65*).

To generate sliding window plots of ChIP-seq signal, bedtools makewindows was used to generate bins of 3kb sliding 300 bp. Bedtools coverage was then used to compute the mean coverage in each bin from the sorted bam files (*66*). Coverage was normalized across samples using RPKM that was calculated as: reads-per-bin/(number of mapped reads (in millions) * bin

length (kb)). Mean value and standard deviation of replicates was then plotted using Python matplotlib.

*RNA-seq data analysis*

Fastq files were aligned to the genome (mm10_synHoxA or mm10_synHoxA_delHoxA custom genomes) using HISAT2 (*67*), using options: -p 20 -q --rna-strandness F. Mapped reads were assigned to annotated genes using the featureCount function in Rsubread (*68*), using options -s 2. Read counts were normalized using the 'rlog' or regularized log transformation in DESeq2 (*69*) and used as inputs for the Principal Component Analysis (PCA). The log2 fold change (FC) and adjusted p-value in gene expression levels between 24h, 48h, and 96h vs. 0h was estimated using DESeq2 and plotted using ggplot (*70*). Normalized counts from DESeq2 were used to generate transgene/endogenous ratio plots, with the counts from endogenous locus being reduced by half to normalize the copy number between *HoxA* clusters. RNAseq track visualization in Fig. S7 was performed using combine tracks tool in IGV (*65*) on bigwigs generated with deeptools bamCoverage (*63*) and subsequent editing in Adobe Illustrator.

To generate the 'housekeeping gene' plots in fig. S8, a coefficient of variance of normalized counts across all samples was calculated for each gene, and the genes were ordered by increasing CV. The normalized counts for the top 10 (least varying) genes for each sample was averaged to generate a "Housekeeping Gene Normalization Factor", and all normalized count values in a sample were divided by this value.

Initial comparison between the parental *HoxA*$^{-/-}$ differentiation RNAseq and all others from this manuscript revealed distinct differences between the mutant line and others. However, after comparison with several other, more recently sequenced but unrelated differentiations from our lab, the difference appears to be correlated with recency of differentiation and sequencing, and not any other known variable (differentiation batch, researcher performing differentiation or library prep, library kit type, sequencing machine, genotype either at the endogenous *HoxA* locus or at *Hprt*). Thus, we used batch correction to eliminate this unrelated variance in the data (*71*), and the batch corrected data from all samples clusters well by differentiation time (fig. S11). ESC and differentiation markers exhibit very similar expression dynamics both before and after batch correction.

*Hi-C data analysis*

Hi-C data was aligned against the custom mm10_synHoxA_delHoxA genome using BWA mem (version 0.7.17) using parameters -M -t 4 and aligning each mate pair independently (*59*). Samtools (version 1.11) was used to sort mapped reads by read name, and the pair_reads.py script in mHiC was used to join mate-pairs into a paired-end SAM file (*64*, *72*). Paired-end read counts were then binned at 10kbp resolution to create the genome-wide contact matrix. The TAD calls displayed in Fig. S18 were produced using HiCseg (version 1.1) (*73*). The HiCseg_linkC_R function in HiCseg was provided with the segment of the 10kbp-resolution contact matrix corresponding to coordinates chrX:43000000-63000000 (mm10_synHoxA_delHoxA custom genome), and the following parameters: nb_change_max=100, distrib="G", model="D". The heatmap was then plotted at coordinates chrX:52400000-53600000 and chrX:52800000-

53300000, with maximum color intensity set at a contact frequency of 40. Non-transgenic mESC data from (*74*) was processed similarly and visualized at coordinates chrX:52400000-53450000 and chrX:52800000-53150000.

Single-molecule Fluorescence In Situ Hybridization

Individual embryoid body samples were embedded in OCT compound (Fisher Healthcare 23730571). OCT blocks were sectioned into 10-micron slices onto Cell-Tak adhesive (Corning 354240) coated coverslips. RNAs were labeled for individual molecule detection using the previously described Third Generation Hybridization Chain Reaction Protocol (*75*). Probes targeting transcripts of interest were designed using custom MATLAB scripts adapted from the MERFISH probe design pipeline (Table S16) (*76*) (https://github.com/timotheelionnet/HCRProbeDesign). Samples were fixed using 4% paraformaldehyde in DPBS (Gibco 14190136), permeabilized using 0.5% Triton X 100 detergent (Bio-Rad 1610407) and subject to HCR probes for hybridization. Probes were allowed to hybridize overnight at 37C. Samples were then washed with 5x SSCT (sodium chloride sodium citrate, Roche 11666681001, plus 0.1% Triton X 100) with decreasing concentrations of formamide and subject to amplification for 3 hours through the addition of fluorescently labeled hairpins (HCR Amplifiers, Molecular Instruments). Amplified samples were washed again using 5x SSCT, stained with DAPI to label DNA, and mounted onto glass slides (Fisherbrand Superfrost 12-550-123) using Prolong Gold anti-fade reagent (Invitrogen P36930).

Images were acquired on a Nikon inverted Ti-E eclipse microscope equipped for wide-field epifluorescence. Fluorescence was excited using one of five lasers (405 nm, 488 nm, 532 nm, 594 nm, or 637 nm), a Nikon 100x Plan Apo NA 1.4 oil immersion objective, and a FLIR Chameleon CM3-U3-31S4M-CS CMOS camera, using the following filter sets (Chroma):

| fluor | dichroic | emission cleanup | emission cleanup 2 |
|---|---|---|---|
| DAPI | ZT405/488/532/640rpc-XT-UF1 | ET425lp | AT460/50m |
| alexa 488 | ZT405/488/532/640rpc-XT-UF1 | ZET488TopNotch | ET520/40m |
| Cy3 | ZT405/488/532/640rpc-XT-UF1 | ZET532TopNotch | ET560/40x |
| alexa 594 | ZT594dcrb-UF1 | ZET594TopNotch | ET625/30m |
| Cy5 | ZT405/488/532/640rpc-XT-UF1 | ZET642TopNotch | ET667/30m |

Maximum intensity projections of each image were created in ImageJ using Z Project. Individual channels of each image were split for downstream processing. DAPI Nuclear masks were generated from the DAPI signal in CellProfiler 4.0.7 with IdentifyPrimaryObjects (Otsu thresholding, 10-50 size range) and dilated to capture the cytoplasmic space of each cell using DilateObjects (disk shape, size 4) (*77*). Individual mRNA transcripts were localized using a previously described spot detection software, Airlocalize (*78*) (https://github.com/timotheelionnet/AIRLOCALIZE), resulting in spot positions and intensities in each field of view. Individual transcripts were assigned to their respective cells using a custom MATLAB script matching each spot to its underlying cell mask based on its spatial coordinates (https://github.com/timotheelionnet/sortSpots). Correlation plots, violin plots and percent positive cells were calculated in MATLAB using custom scripts. Number of cells analyzed per condition and cell line is included in Table S17.

Antibody staining

EBs were collected by centrifugation at 100 rcf, washed once in PBS, washed once in cold PBS, then resuspended in 500 µL cold 4% PFA and incubated on ice for 10 min. EBs were then washed 3x in cold PBS, removing all but ~100 µL after the last wash. 500 µL 30% sucrose was added and EBs were allowed to equilibrate on ice until they settled to the bottom of the tube. All but ~100 µL sucrose/PBS was removed and ~200 µL OCT Compound (Tissue-Tek, 4583) was added and mixed with a p200 pipette tip. EBs were spun down at 100 rcf and gently pipetted in a column in a peel-away (Polysciences, 18646A) filled with OCT compound. OCT block was frozen in a bath of dry ice and ethanol. 12 µM thick slices were cut from block and fixed to a slide. Slide was blocked and permeabilized in 1x PBS, 0.2% Triton X-100, 5% BSA for 30 min. at room temp. Blocking buffer was removed and replaced with 100 µL 1x PBS, 0.1% Triton X-100, 5% BSA containing a 1:500 dilution of rabbit anti-Hoxa5 antibody (CU1026, gift from Jeremy Dasen) and 1:100 dilution of mouse anti-Isl1/2 antibody (DSHB 39.4D5) and incubated overnight at 4ºC. Slides were washed 3x in 1x PBS, 0.1% Triton X-100 then incubated 1 hr. at room temp. with a 1:1000 anti-Rb-Alexa-568 2º antibody (Invitrogen, A11036) and 1:1000 anti-Ms-Alexa-488 2º antibody (Invitrogen, A21202). Slides were washed 3x in 1x PBS, 0.1% Triton X-100 then mounted in Fluoroshield with DAPI (Sigma, F6057), covered and sealed in a coverslip, and imaged using a Nikon Eclipse Ti inverted fluorescence microscope.

Images were processed using a custom ImageJ macro running on FIJI (*79*). Briefly, multichannel images were split, z-projected, background subtracted, nuclei detected, and fluorescence intensity of the DAPI and Hoxa5 channels measured in each detected nucleus. The mean intensity values were saved as a csv file and opened in R for data normalization and plotting. Mean intensity of HoxA5 in each nucleus was divided by the mean intensity of DAPI, then samples were normalized to $HoxA^{-/-}$ and $HoxA^{+/+}$ by subtracting the median $HoxA^{-/-}$ value and dividing by the $HoxA^{+/+}$ from each intensity value. To correct for copy number, intensity values for *SynHoxA* lines were multiplied by two. Box and jitterplots were generated using normalized data with ggplot2 (*70*) and p-values were generated using ggpubr v.0.4.0 (https://CRAN.R-project.org/package=ggpubr). Number of cells analyzed per condition and cell line is included in Table S17.

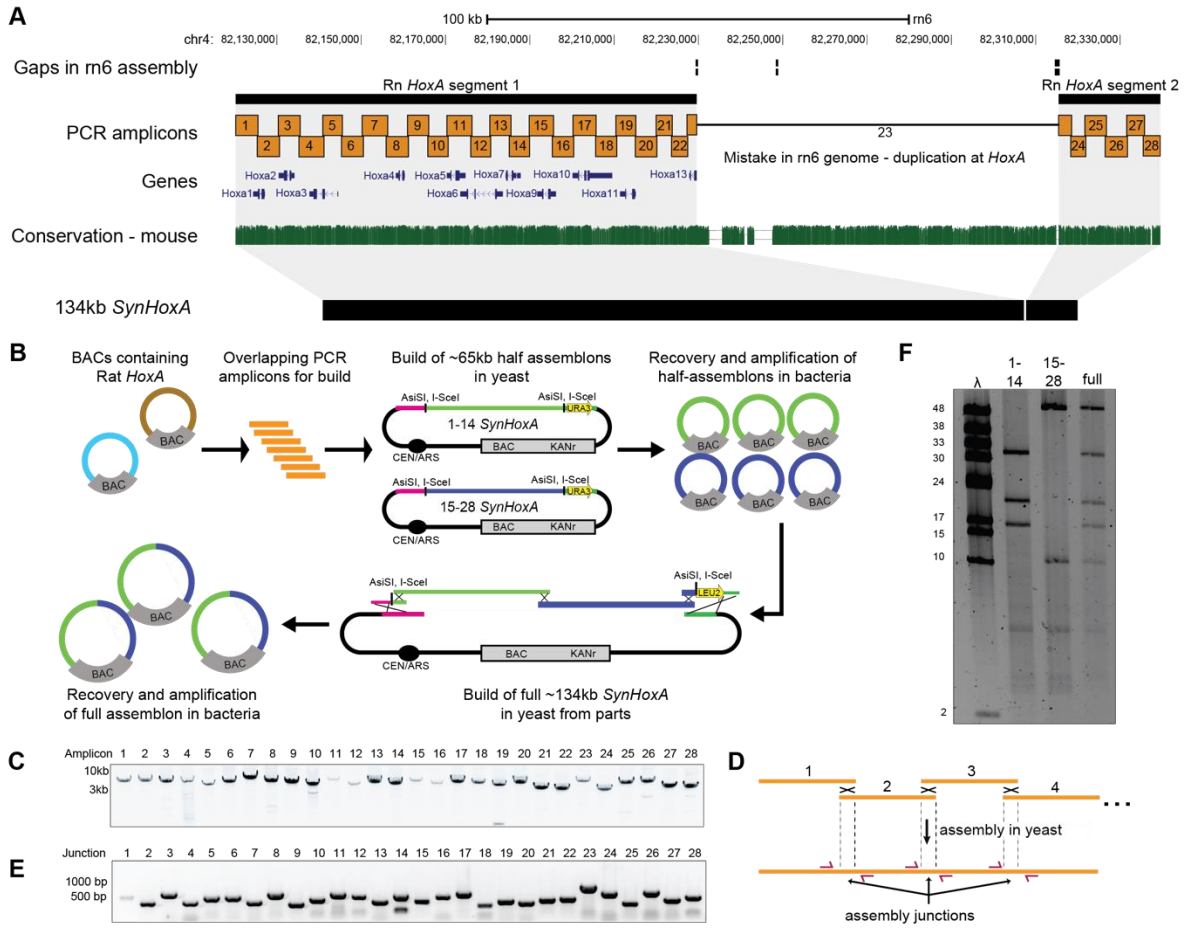


**Figure S1: Build of 134kb *SynHoxA* assemblon.**

(A) Layout of rat *HoxA* locus in the rn6 genome assembly. The rn6 genome includes an erroneous duplication at the *HoxA* locus between gaps in the assembly. The *SynHoxA* assemblon sequence is based on bringing together the two 'separate' RnHoxA segments. The sequence was segmented into 28 ~5kb PCR amplicons with terminal homology of ~200bp to adjacent amplicons. Conservation to the mouse genome is depicted using the multiz track from the UCSC genome browser. (B) Schematic depicting the assembly workflow for the 134kb *SynHoxA* assemblon. BACs containing Rat *HoxA* were used as PCR template to generate 28 segments tiling the entire *HoxA* locus. These segments were co-transformed into yeast with appropriate linkers and an assembly vector to build two ~65kb half assemblons into centromeric yeast-bacteria shuttle vectors. These half assemblons are recovered to bacteria and amplified. Full 134kb assemblon was built from half assemblons after releasing them from the vector using a

terminal restriction enzyme (*AsiSI*) and transforming into yeast. Full assemblon was then recovered from yeast into bacteria for amplification and verification. (C) Agarose gel of the 28 PCR amplicons that tile the 134kb *SynHoxA* assemblon. (D) Strategy to PCR-screen yeast colonies derived from assembly experiments. Primers (red arrows) span assembly junctions and test presence/absence of amplicons in many yeast colonies. Reproduced from Mitchell et. al, 2021, with permission from authors. (E) Agarose gel showing one yeast colony carrying the full 134kb *SynHoxA* assemblon verified manually for the presence of all assembly junctions, using the strategy outlined in panel D. (F) Half and Full 134kb *SynHoxA* assemblon BACs purified from *E.coli* were digested with PvuI and separated using field inversion gel electrophoresis (FIGE). Lambda monocut ladder sizes are indicated in kb. Band sizes correspond to expected fragments.

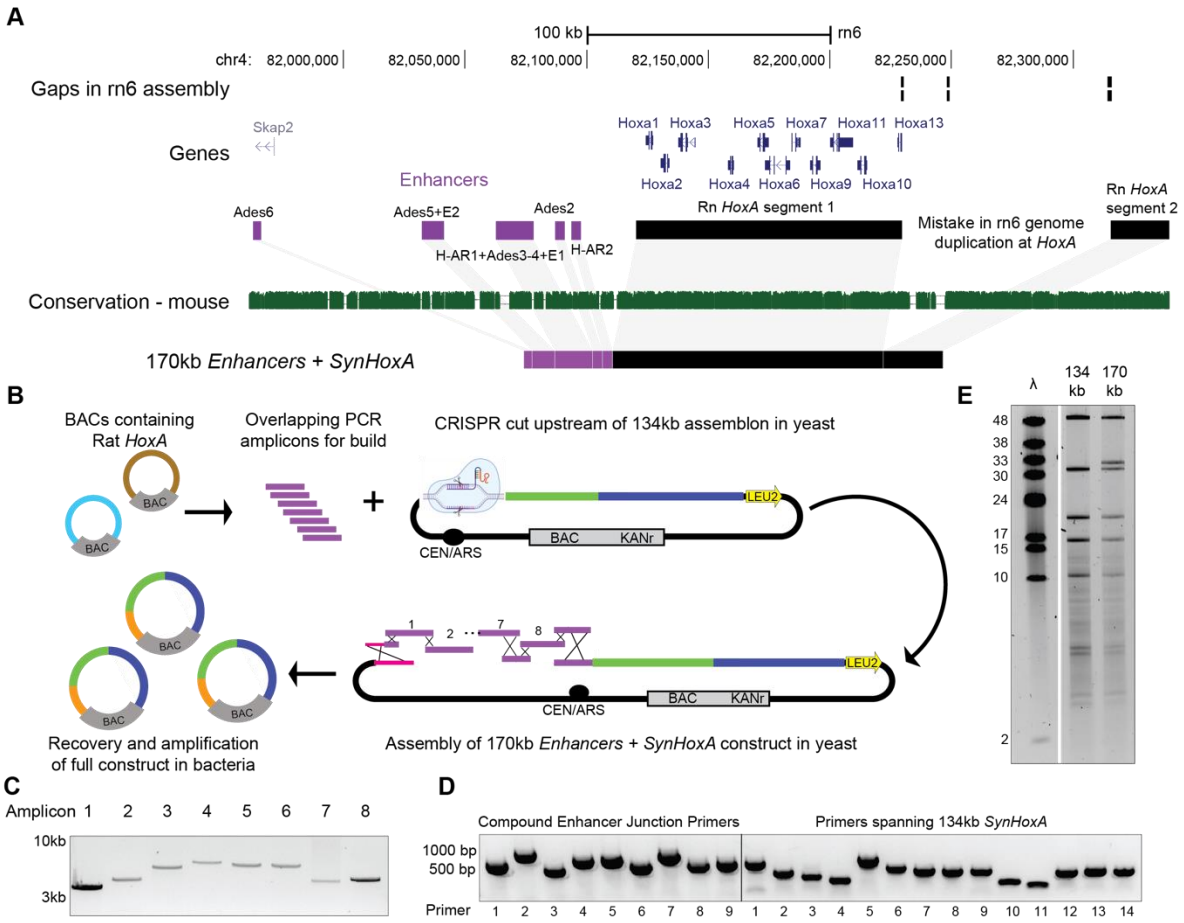


**Figure S2: Build of 170kb *Enhancers+SynHoxA* assemblon.**

(A) Layout of rat *HoxA* locus from the rn6 genome assembly depicting genes, Rn *HoxA* cluster segments in black and previously identified distal enhancers in purple. The *Enhancers+SynHoxA* assemblon sequence is made by stringing all the enhancers directly upstream of the *SynHoxA* assemblon sequence. Conservation to mouse genome is depicted using multiz track from the UCSC genome browser. Ades is short for *HoxA* developmental early side. (B) PCR amplicons tiling enhancer sequences were generated from Rat *HoxA* BACs and co-transformed into a yeast strain containing the 134kb *SynHoxA* assemblon with a gRNA vector targeting the left terminus of the 134kb assemblon. The enhancer PCR amplicons were used to repair this break, resulting in the construction of the 170kb *Enhancers+SynHoxA* assemblon. Assemblon was recovered into bacteria for amplification and verification. (C) Agarose gel of the 8 PCR amplicons containing enhancer sequences. (D) Agarose gel showing one yeast colony tested for the presence of novel enhancer assembly junctions and with primers spanning 134kb *SynHoxA*. (E) 134kb and 170kb

assemblon BACs purified from *E.coli* were digested with *PvuI* and separated using FIGE. Lambda monocut ladder sizes are indicated in kb. Band sizes correspond to expected fragments.

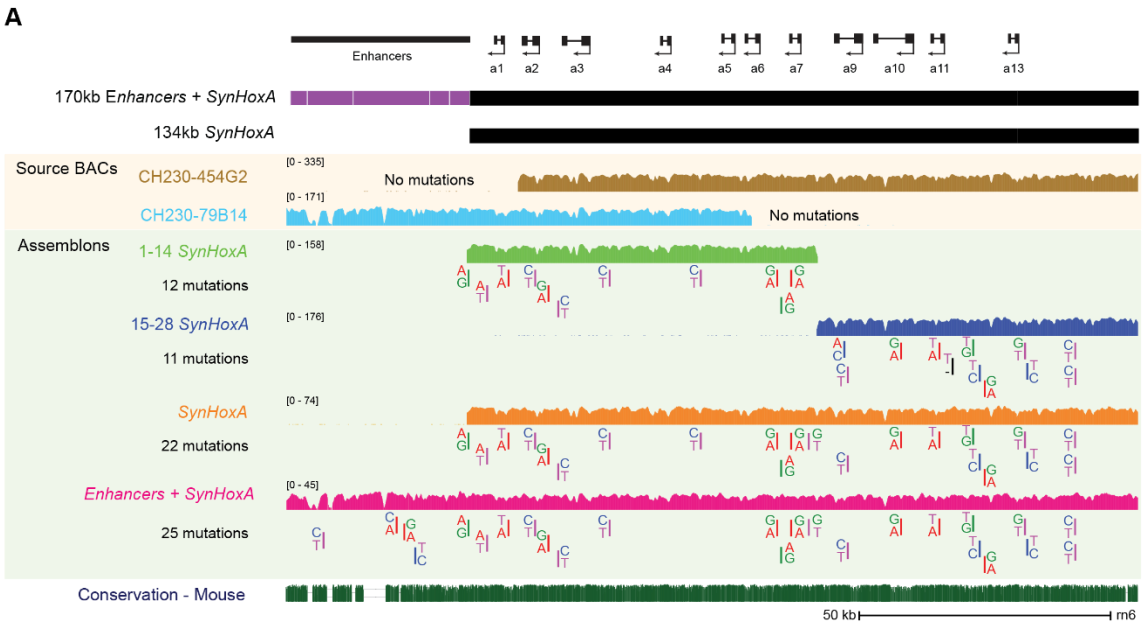


**Figure S3: Mutations in *SynHoxA* assemblon that were generated during construction.** Sequencing data from various stages of assemblon construction: source rat BACs, half-assemblon BACs and full-assemblon BACs. Below each coverage track, variant positions in comparison to the reference rn6 genome are depicted with reference allele (top) and variant allele (bottom). Data shown here are aligned to the rat reference genome rn6 and smoothed over 500bp.

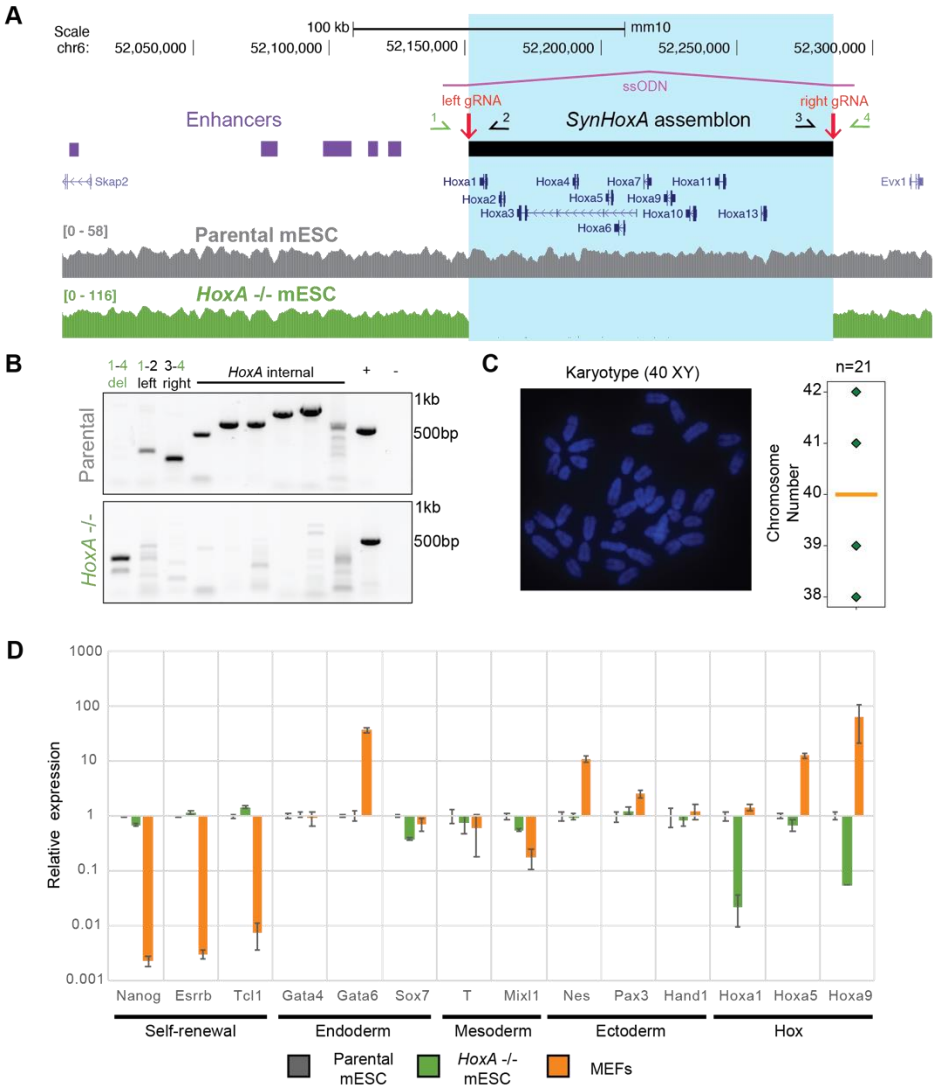


**Figure S4: Generation of *HoxA*$^{-/-}$ mESCs by CRISPR/Cas9 induced deletion.**

(A) Top, layout of mouse *HoxA* locus with sequence corresponding to the 134kb *SynHoxA* assemblon marked with a black box and enhancers depicted with purple boxes. Deletion was induced by targeting with two guide RNAs (Left and Right) and by providing a single stranded oligo donor (ssODN) bridging the deletion. Primers used for genotyping PCR are depicted (1-4). Bottom, whole genome sequencing data for parental and *HoxA*$^{-/-}$ mESCs aligned to mouse reference genome mm10 and smoothed over 2000bp. (B) Genotyping PCR for verifying *HoxA*$^{-/-}$ mESCs in comparison to parental cells using deletion specific primer pairs depicted in (A) and primers internal to the deletion. +, positive control with primers amplifying mouse *Furin* locus; –, no primer control. (C) Karyotyping of *HoxA*$^{-/-}$ mESCs by metaphase spreads to confirm euploidy. One representative spread is shown. Quantification of 21 spreads is presented on the

right as a boxplot, confirming that the cells have an euploid mean chromosome number of 40. (D) qRT-PCR data from parental mESCs, *HoxA*$^{-/-}$ mESCs and Mouse Embryonic Fibroblasts (MEFs) for a range of genes is shown. Relative expression is presented as normalized to *Gapdh* and parental mESCs. Error bars indicate standard deviation from 4 technical replicates.

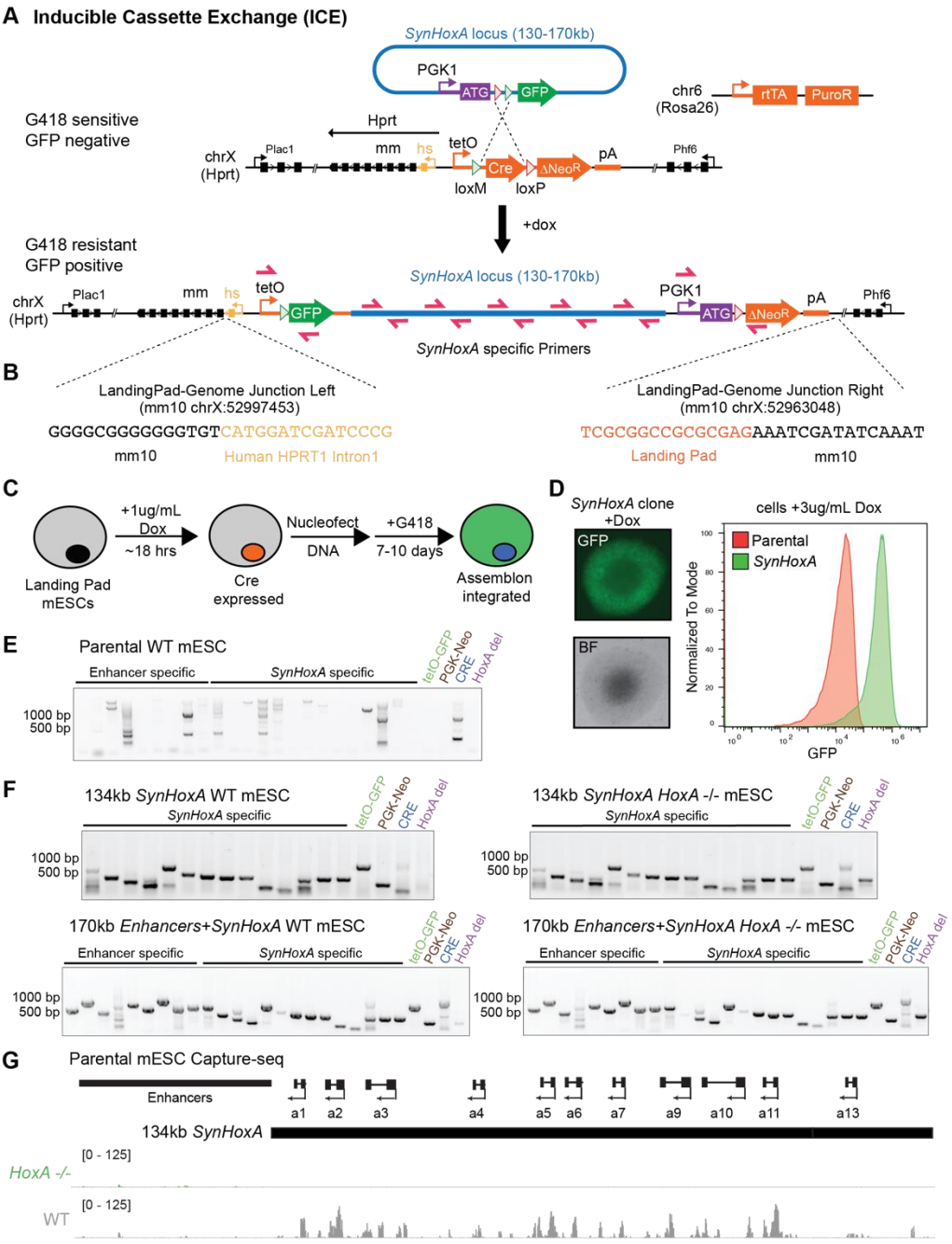


**Figure S5: Delivery and PCR genotyping of *SynHoxA* variant assemblons.**

(A) Schematic of Inducible Cassette Exchange (ICE) for site-specific delivery of assemblons to mESCs. A resident landing pad, integrated at the mouse *Hprt* locus, contains a *Cre* recombinase gene, driven by a tetracycline inducible promoter (TRE) and is flanked by heterotypic loxM and loxP sites. A promoter-less Neomycin resistance gene lacking a start codon (ΔNeo) is found downstream of *Cre*. The reverse tetracycline transactivator (rtTA) is expressed from the *Rosa26* locus. The assemblon vector contains a delivery cassette (PGK1-ATG-loxP-loxM-GFP). During

cassette exchange, two Cre mediated recombination events result in the placement of GFP under the control of the tetO promoter, donation of the *PGK1* promoter and ATG start codon to ΔNeo, as well as loss of the *Cre* gene. This gives rise to G418-resistant (the Neo gene confers resistance to G418, also known as Geneticin), GFP positive cells. Primers used for genotyping are indicated in red. (B) Sequence and coordinates of landing pad junctions with the mouse genome. (C) mESCs bearing the ICE landing pad on the X chromosome are treated with 1 μg/mL Doxycycline to induce *Cre* expression. DNA is nucleofected and cells are selected with G418 for 7-10 days until clones are grown out. Clones are then picked for genotyping and sequencing. (D) Left, an image of a representative *SynHoxA* clone is shown. Right, flow cytometry data from parental cells (red) and *SynHoxA* cells (green) after treatment with 3 μg/mL doxycycline. (E-F) Agarose gels showing genotyping of parental WT mESCs (E) and from clones arising from delivery of 134kb *SynHoxA* and 170kb *Enhancers+SynHoxA* to WT and *HoxA*$^{-/-}$ mESCs (F). Clones were screened using *SynHoxA*-specific primers that span the length of the assembly, primers that span novel junctions formed with the genome (tetO-GFP, PGK-Neo) and primers that confirm overwriting of the *Cre* gene. In addition, the presence or absence of the endogenous *HoxA* cluster deletion was confirmed by deletion-specific primers (see Fig. S4). (G) Capture-sequencing data generated from parental WT and *HoxA*$^{-/-}$ mESCs using rat *HoxA* sequence as bait. Data are aligned to custom mouse reference genomes and normalized for sequencing depth. There is minimal cross-mapping between endogenous *HoxA* and *SynHoxA* loci.

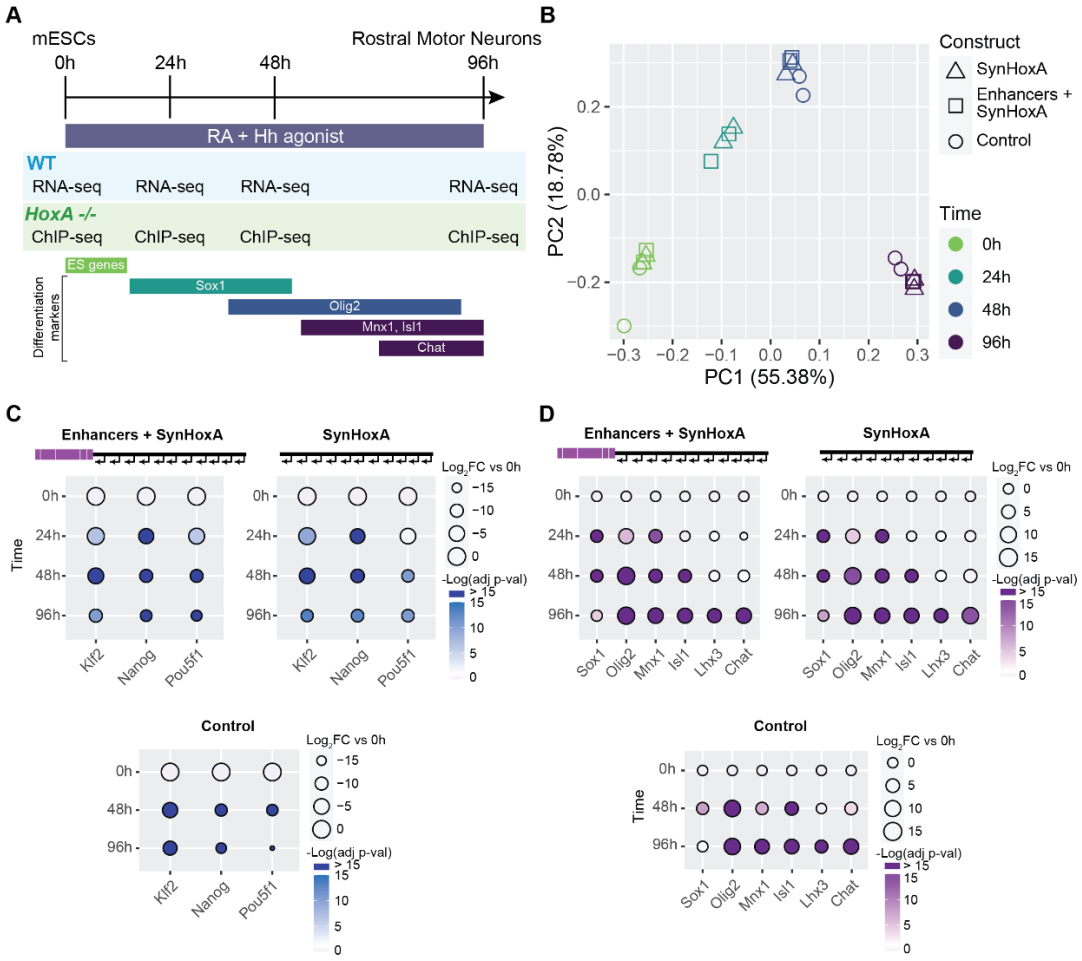


**Figure S6: *SynHoxA* variants respond to patterning signals appropriately during *in vitro* spinal cord differentiation.**

(A) Overview of *in vitro* differentiation protocol. In response to RA and Hedgehog patterning signals, ESCs differentiate into MNs by transitioning through key progenitor states. WT and *HoxA*$^{-/-}$ mESCs harboring *SynHoxA* assemblons were analyzed by RNA-seq and ChIP-seq at indicated time points. (B) Principal Component Analysis (PCA) of *SynHoxA* variant and control parental WT mESC lines RNA-seq datasets reveals clustering largely by time during the differentiation protocol (each data point represents independent differentiations). (C) Log2 fold change of pluripotency marker genes from RNA-seq data (n=2). *SynHoxA* variants downregulated pluripotency markers during differentiation as expected (n=2). (D) Log2 fold change of MN differentiation markers from RNA-seq data (n=2). *SynHoxA* variants upregulated differentiation markers as expected.

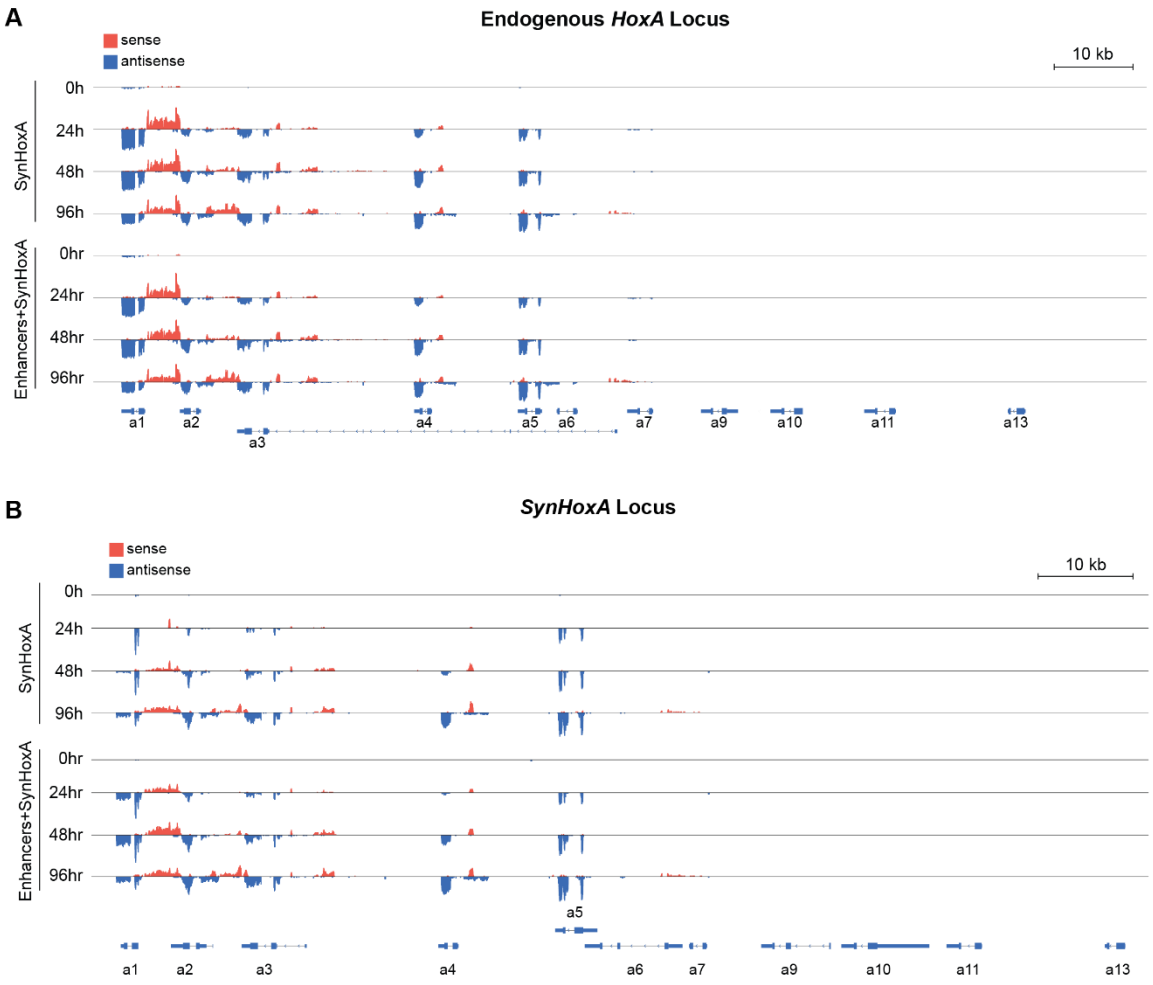


**Figure S7: Visualization of raw gene expression data through differentiation.**

RNA-seq data through RA induced differentiation of *SynHoxA* and *Enhancers+SynHoxA* lines with an intact endogenous *HoxA* cluster are presented. Reads mapping to the endogenous cluster are shown in (A) and reads mapping to the *SynHoxA* cluster are in (B). Reads mapping to the sense strand are in red and antisense reads are in blue. Expression of both coding and non-coding transcripts was observed.

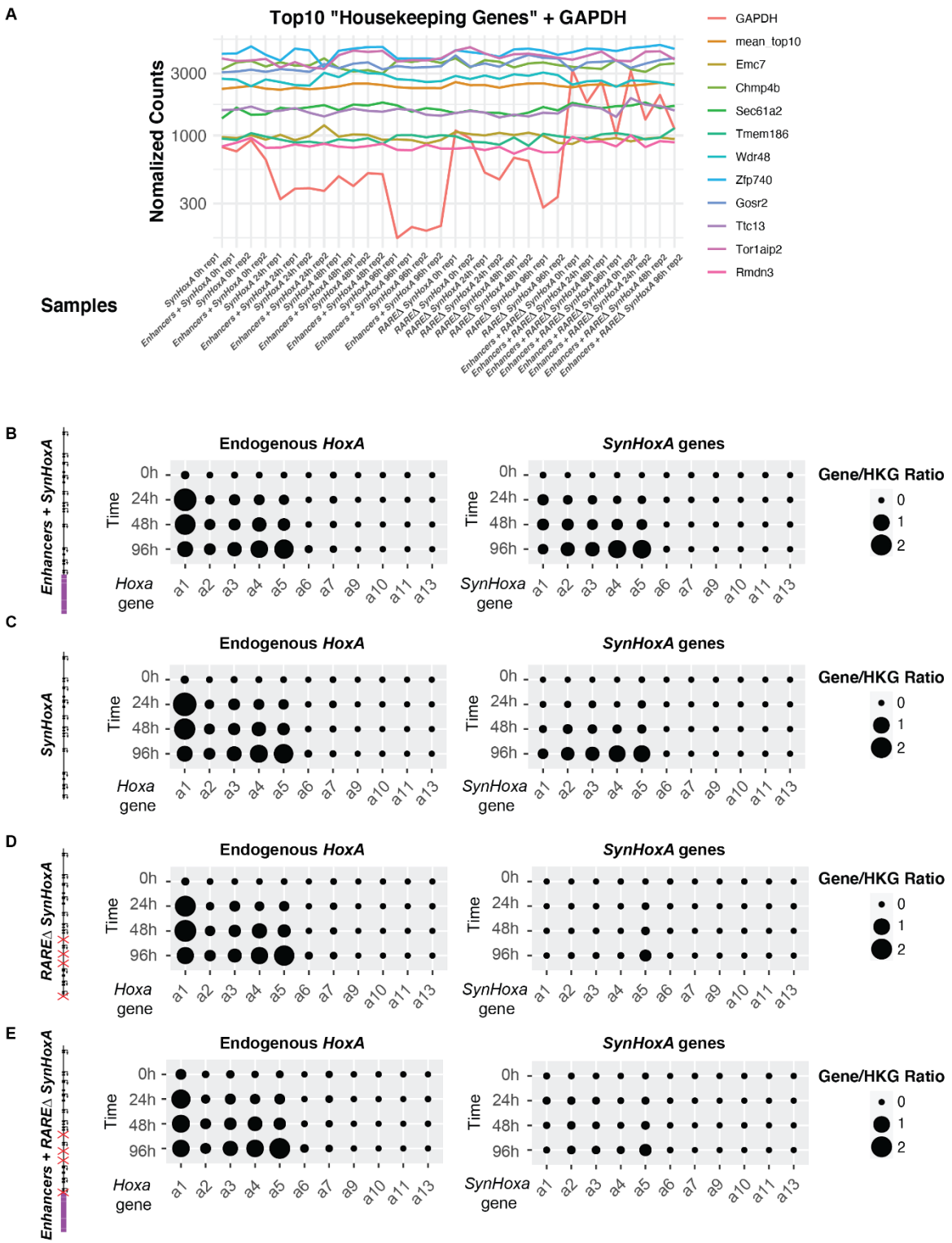


**Figure S8: Normalization of RNA-seq data to invariant housekeeping genes.**

(A) DESeq2 normalized counts of the top 10 most 'invariant' genes across all samples in the RNA-seq data. All endogenous *Hoxa* and *SynHoxa* gene counts were then normalized to the

mean of these 10 housekeeping (HKG) genes (n=2). (B) HKG normalized counts from *Enhancers+SynHoxA.* (C) HKG normalized counts from *SynHoxA.* (D) HKG normalized counts from *RARE*Δ *SynHoxA.* (E) HKG normalized counts from *Enhancers+RARE*Δ *SynHoxA.*

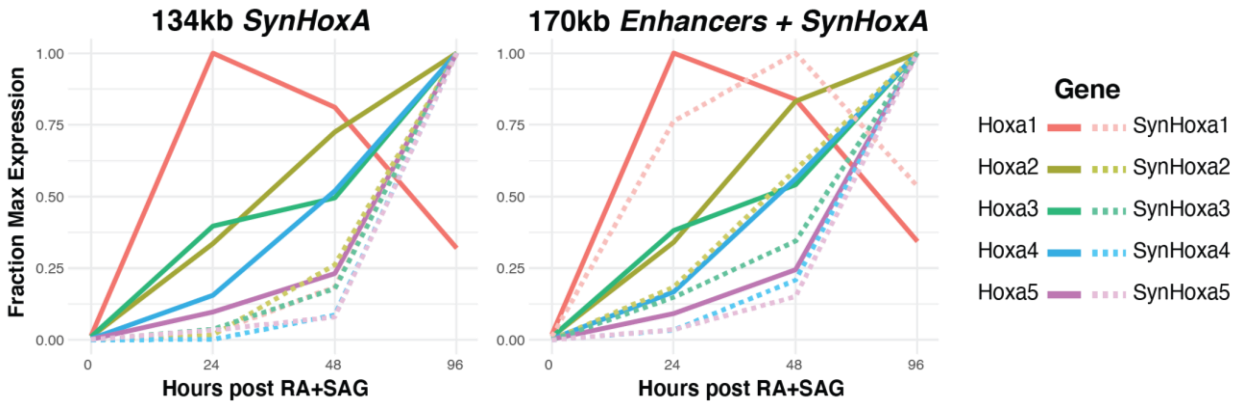


**Figure S9: Colinearity plots of endogenous and synthetic *Hoxa* genes.** Expression of each gene from RNA-seq is presented as a fraction of the maximum expression level of that gene over the course of the differentiation protocol. Genes from the endogenous *HoxA* cluster are depicted as solid lines whereas genes from the ectopic *SynHoxA* cluster are depicted as dotted lines.

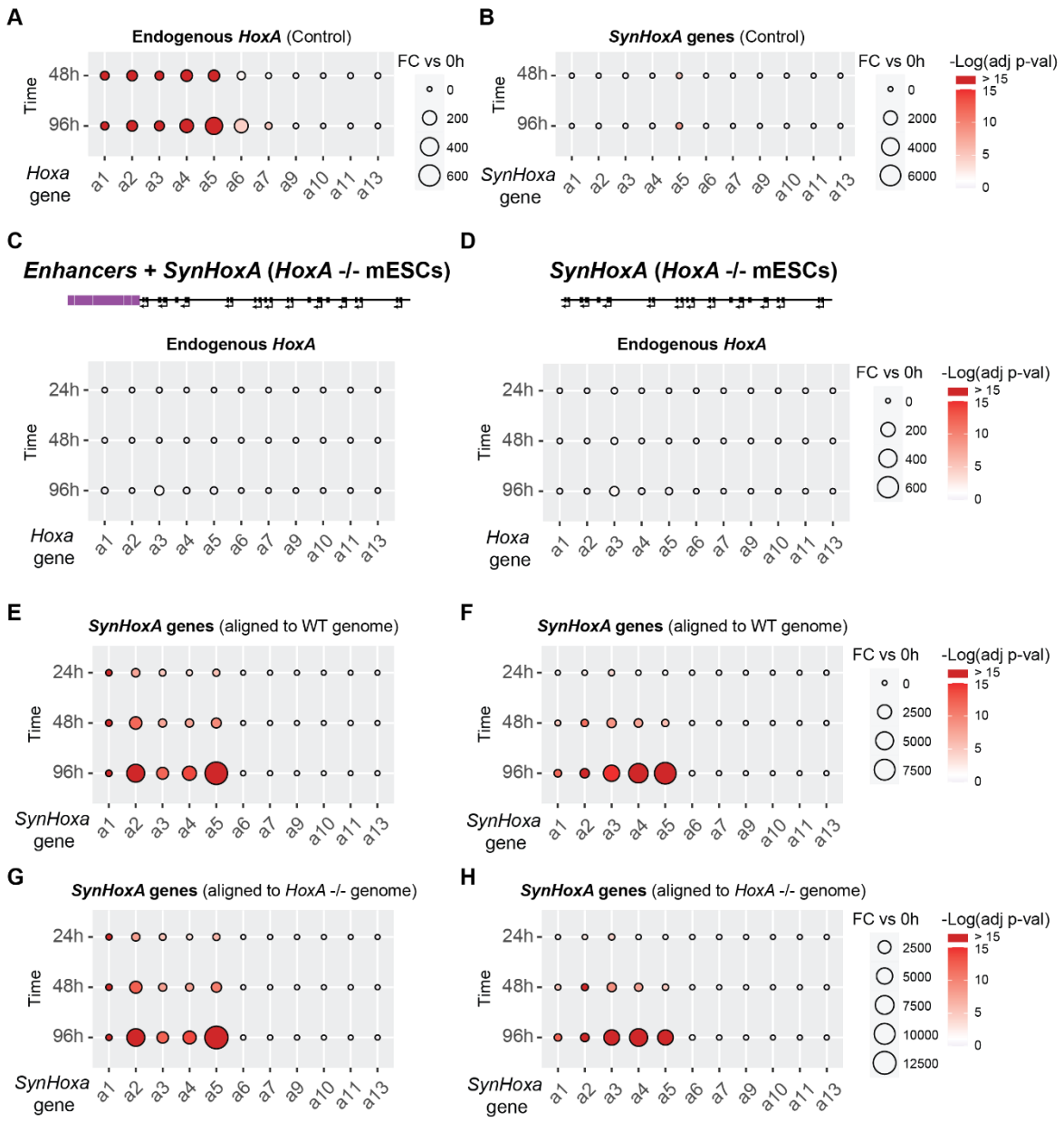


**Figure S10: No mapping issues are revealed in the RNA-seq analysis.**

(A-B) The fold change of *SynHoxA* or endogenous mouse *HoxA* genes from RNA-seq data during differentiation in control lines, which do not contain *SynHoxA* variant clusters (n=2). The endogenous *HoxA* cluster induces the correct set of genes, and has some weak *Hoxa6* expression at the latest time point. Only a minor amount of reads map to *SynHoxa5* at 96h in control lines with no *SynHoxA* variant clusters. RNA-seq data are aligned to a modified mm10 mouse genome which contains the *SynHoxA* sequence inserted at the *Hprt* locus. Only uniquely mapped reads were kept and used in downstream analysis. (C-H) mapping tests similar to that in (A) for lines

lacking the endogenous *HoxA* cluster but containing *Enhancers+SynHoxA* (C, E, G) and *SynHoxA* (D, F, H) at *Hprt.* (C-D) No mapping was observed to endogenous *HoxA* genes in cell lines lacking the endogenous *HoxA* cluster as expected. (E-H) Mapping RNA-seq data to a genome either containing (WT, E-F) or lacking the endogenous *HoxA* cluster (*HoxA* $^{-/-}$, G-H) did not affect the analysis.

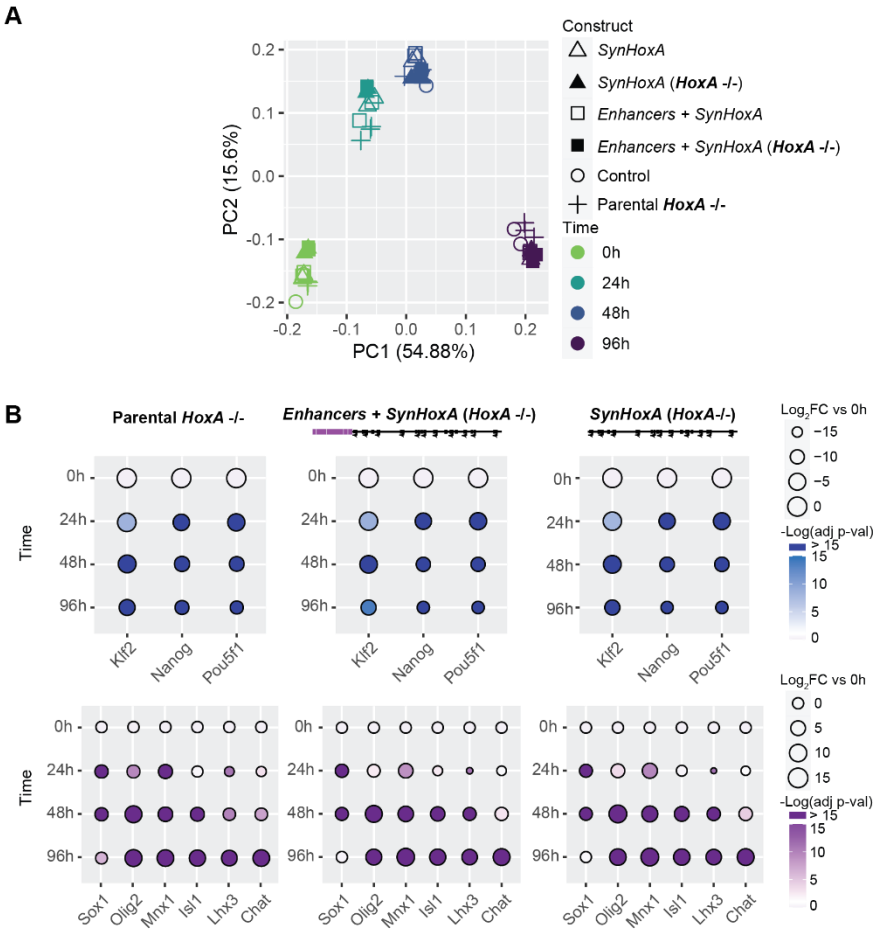


**Figure S11: mESCs with *SynHoxA* assemblons in the *HoxA*$^{-/-}$ background differentiate well into MNs.**

(A) Principal Component Analysis (PCA) of batch corrected RNA-seq datasets reveals clustering largely by time during the differentiation protocol (each data point represents independent differentiations). This is true regardless of genetic background i.e. WT or *HoxA*$^{-/-}$ and ectopic *SynHoxA* variants that are integrated (see Methods). (C) The log2 fold change of pluripotency markers and MN differentiation genes from RNA-seq data (n=2). Pluripotency markers were downregulated and MN markers were upregulated during differentiation as expected for *SynHoxA* mESCs lacking the mouse *HoxA* cluster.

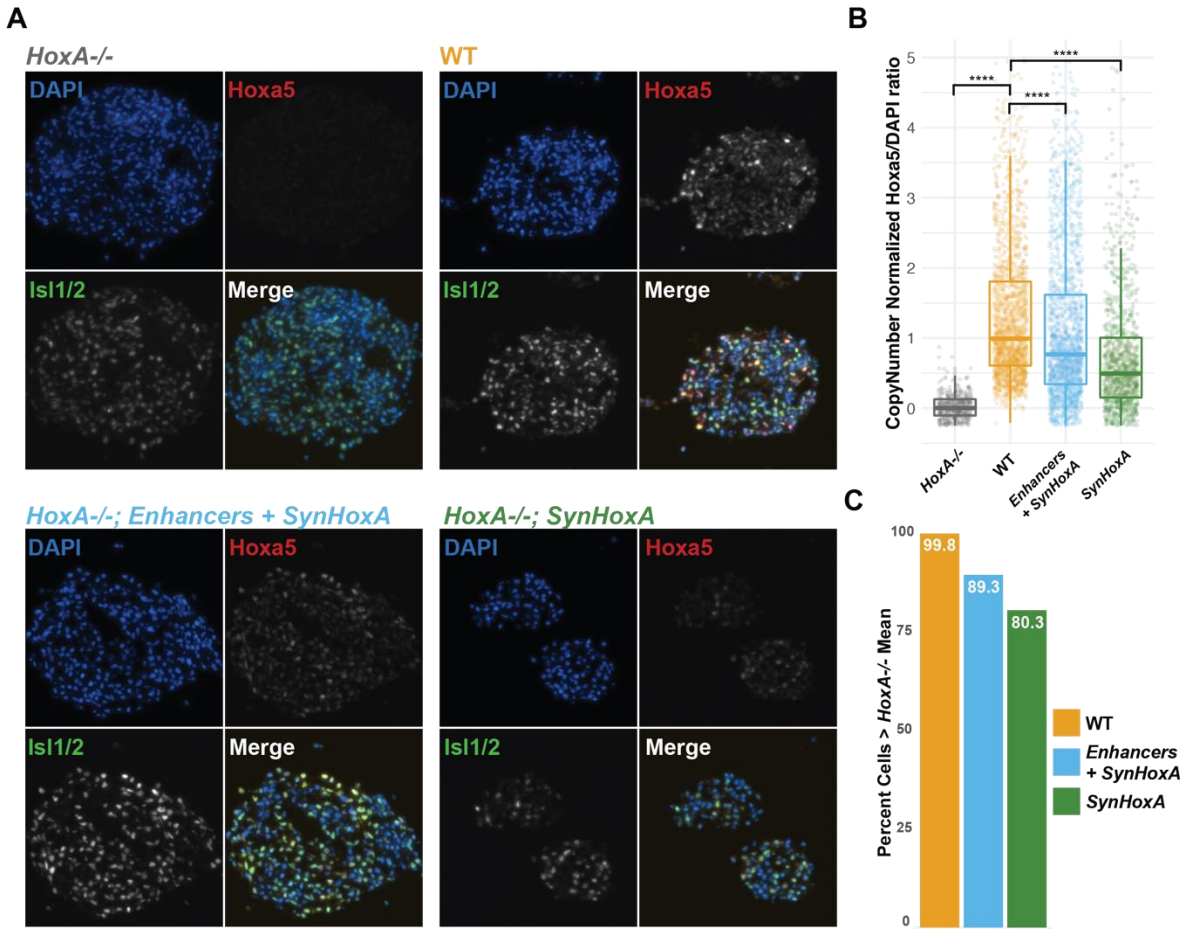


**Figure S12: Hoxa5 antibody staining reveals that the number of Hoxa5 positive cells does not change dramatically in response to RA across *SynHoxA* lines.**

(A) Representative images of Hoxa5, Isl1/2 (MN marker), DAPI (nuclear marker) in WT, *HoxA*$^{-/-}$, *SynHoxA* (*HoxA*$^{-/-}$,) and *Enhancers+SynHoxA* (*HoxA*$^{-/-}$) lines at 96h post RA. (B) Boxplot of Hoxa5 levels across the 4 lines relative to DAPI, normalized to the number of *HoxA* alleles present in each. Dark horizontal line indicates the mean. '****' indicates adjusted p-value <2e$^{-16}$ (C) Percent of cells whose Hoxa5 level is greater than the mean in the *HoxA*$^{-/-}$ line. Numbers of analyzed cells per line is in Table S17.

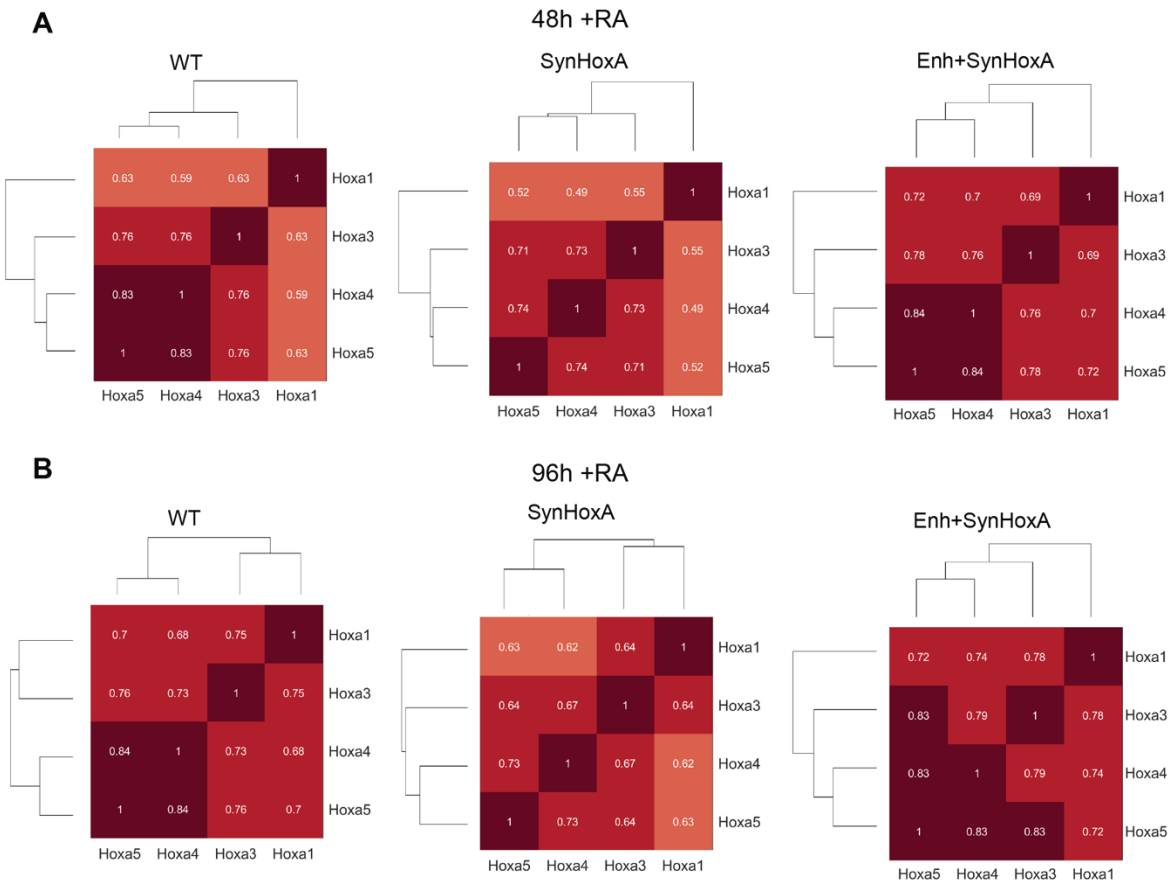


**Figure S13: The expression of any anterior *Hoxa* gene is highly correlated with the expression of the other genes in response to RA at the single cell level.**

Heatmap depicting Pearson's correlation coefficients calculated from transcripts per cell from RNA single-molecule Fluorescence In Situ Hybridization (RNA smFISH) measurements of *Hoxa1*, *Hoxa3*, *Hoxa4* and *Hoxa5* at 48h (A) and 96h (B) after exposure to RA in WT, *Enhancers+SynHoxA* (*HoxA*$^{-/-}$) and *SynHoxA* (*HoxA*$^{-/-}$) lines. Number of analyzed cells per condition is in Table S17.

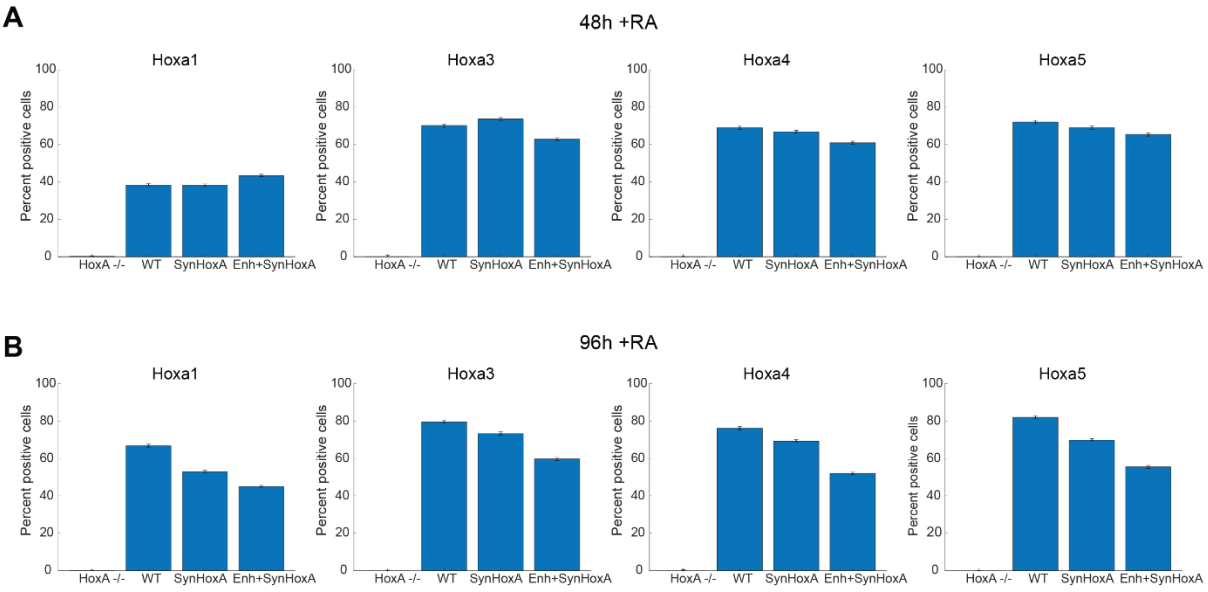


**Figure S14: Number of cells positive for anterior *Hoxa* genes does not change dramatically in response to RA across *SynHoxA* lines.**

Barplots of percent positive cells for *Hoxa1*, *Hoxa3*, *Hoxa4* and *Hoxa5* at 48h (A) and 96h (B) after exposure to RA in *HoxA*$^{-/-}$, WT, *Enhancers+SynHoxA* (*HoxA*$^{-/-}$) and *SynHoxA* (*HoxA*$^{-/-}$) lines as measured by RNA smFISH. Cells with at least one transcript were considered positive. Error bars represent standard error. Number of cells analyzed per condition is in Table S17.

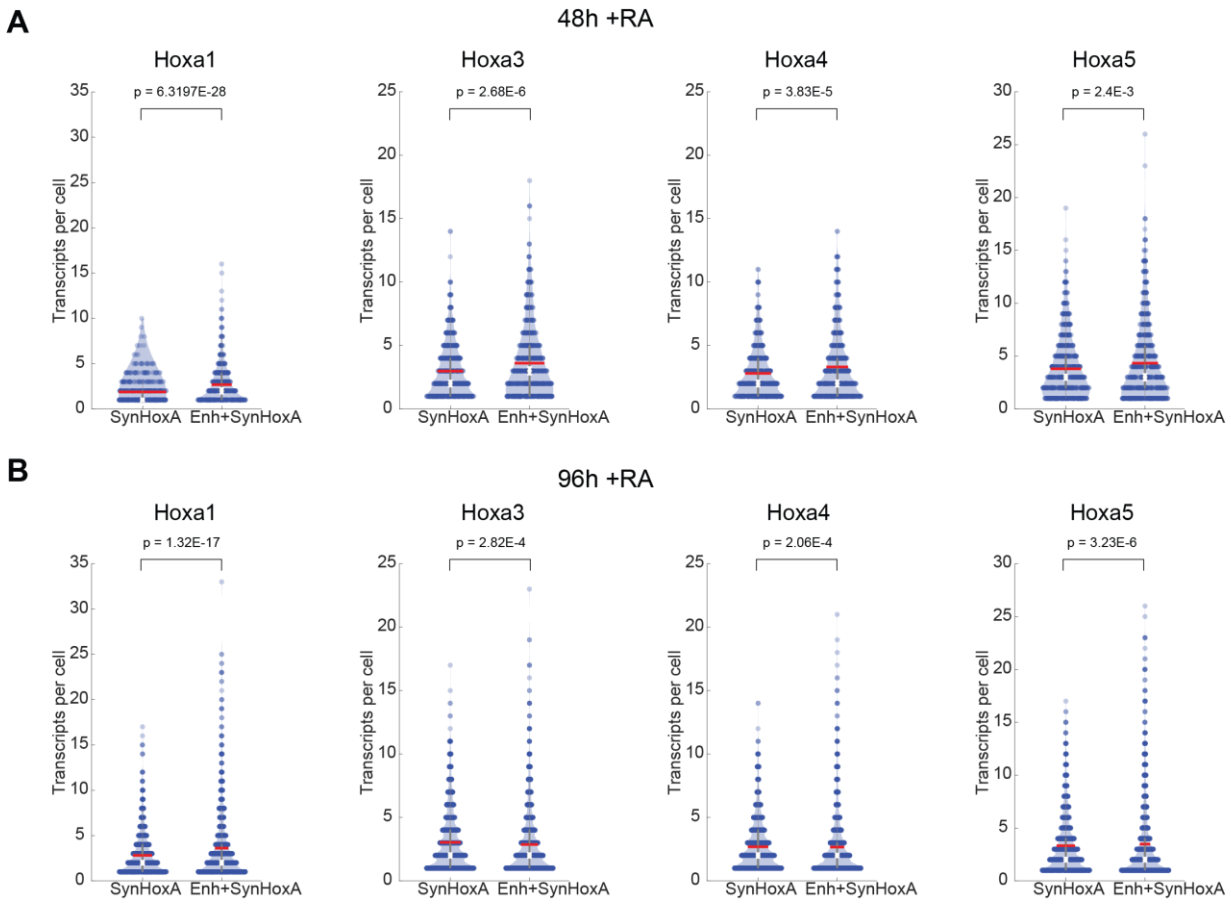


**Figure S15: Distal enhancers increase transcriptional output at the single-cell level.**
Violin plots of *Hoxa1*, *Hoxa3*, *Hoxa4* and *Hoxa5* transcripts per positive cell in *Enhancers+SynHoxA* (*HoxA*$^{-/-}$) and *SynHoxA* (*HoxA*$^{-/-}$) lines at 48h (A) and 96h (B) after exposure to RA as measured by RNA smFISH. Horizontal red line indicates the mean. Indicated p-values are derived from a two-sample t-test. Number of cells analyzed per condition is in Table S17.

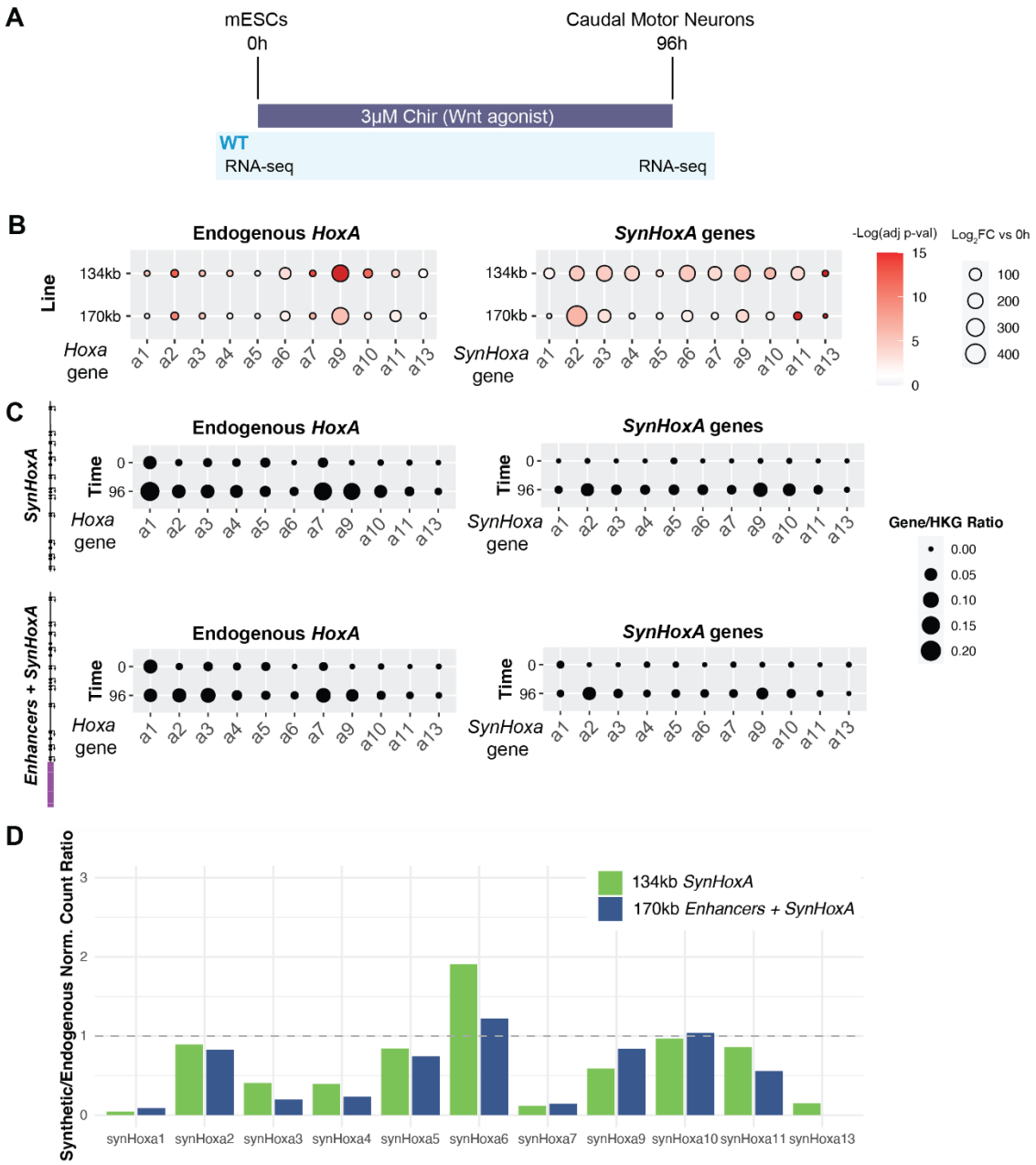


**Figure S16: Ectopic *Hox* clusters are able to respond appropriately to a posterior Wnt signal.**

(A) Overview of *in vitro* differentiation protocol. A Wnt agonist (Chir) was added to undifferentiated cells bearing *SynHoxA* and *Enhancers+SynHoxA*. Samples were collected at 96h post exposure to Chir. (B) Log2 fold change relative to 0h of endogenous mouse *HoxA* or *SynHoxA* genes from RNA-seq data during Chir differentiation in the *SynHoxA* and *Enhancers+SynHoxA* lines. Shade of individual bubbles indicates adjusted p-value. (n=2) (C) Read counts from endogenous and *SynHoxA* genes in the *SynHoxA* and *Enhancers+SynHoxA*

lines were normalized to top 10 most invariant “housekeeping” (HKG) genes in the RNA-seq data. (n=2) (D) Ratios of gene expression for *SynHoxA* genes to endogenous mouse *HoxA* genes in the *SynHoxA* and *Enhancers+SynHoxA* lines at the 96h time point. (n=2). Counts for the endogenous *HoxA* genes were halved to normalize for two endogenous *HoxA* vs. one ectopic *SynHoxA* copy.

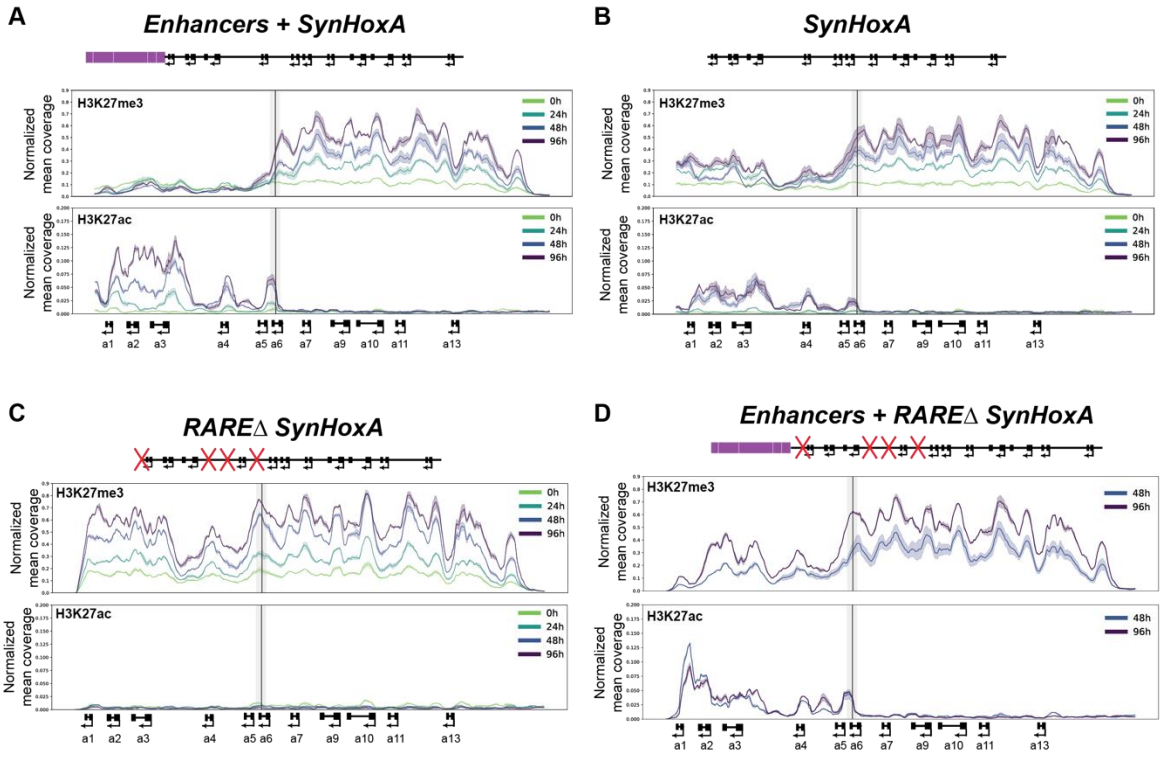


**Figure S17: H3K27me3 and H3K27Ac distribution across *SynHoxA* during MN differentiation.**

RPKM normalized mean coverage on 3kb windows sliding 300bp across *SynHoxA* of H3K27me3 and H3K27Ac ChIP-seq data for each *SynHoxA* cell line. Solid line follows the mean of two replicates with the shaded area indicating the standard deviation. Data are aligned to a custom reference genome (see Methods).

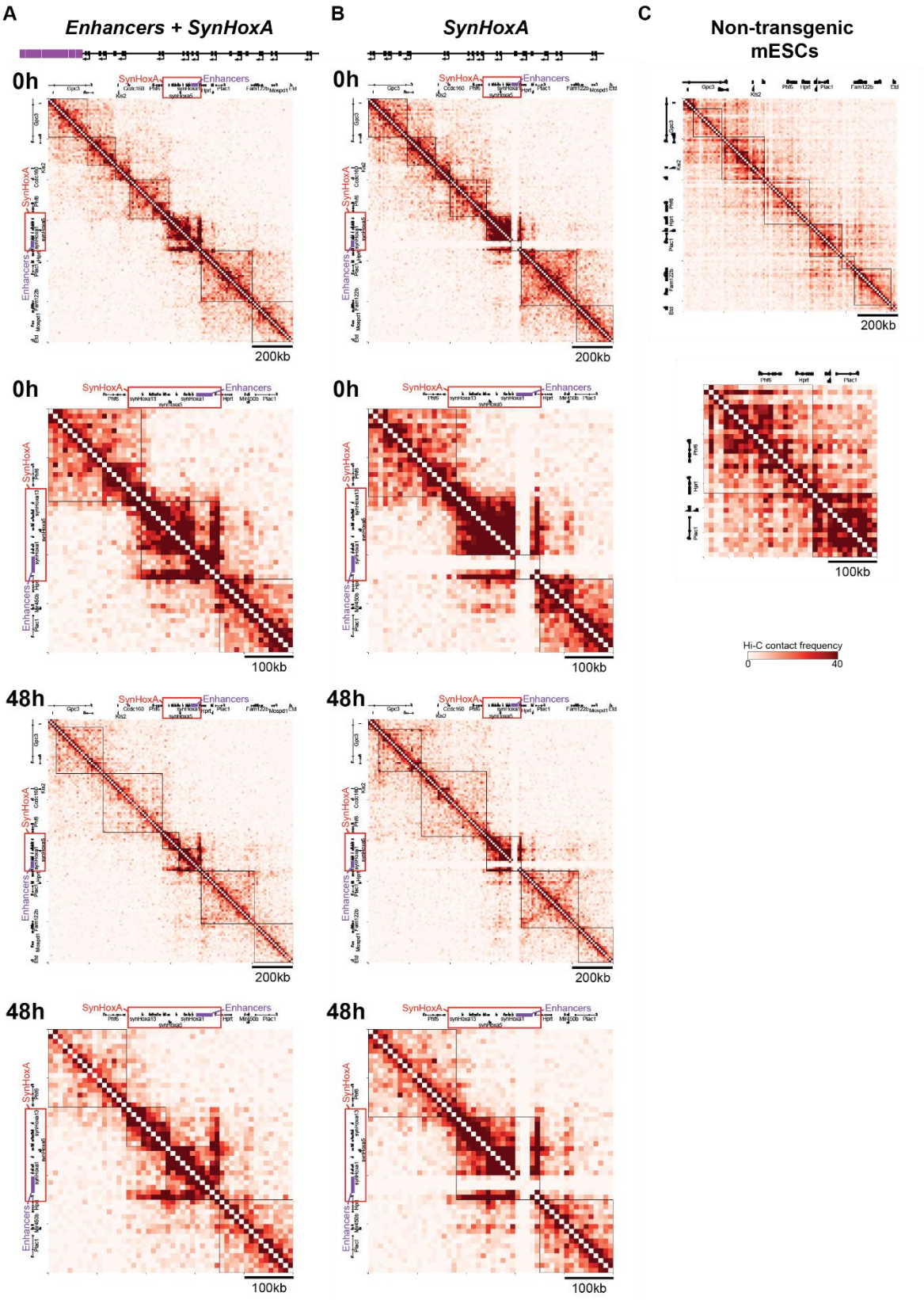


**Figure S18: Ectopic *SynHoxA* clusters self-organize in 3D during differentiation.**

Heatmaps of Hi-C data during MN differentiation from mESCs lacking the endogenous *HoxA* cluster that contain either *Enhancers+SynHoxA* (A), *SynHoxA* (B) or no transgene (C) at *Hprt*. Black lines indicate topological boundaries called by an unbiased algorithm, HiCSeg. A topological boundary formed between *SynHoxa5* and *SynHoxa6* in *Enhancers+SynHoxA*, mirroring endogenous organization. Control non-transgene HiC data is sourced from Dixon et. al 2012.

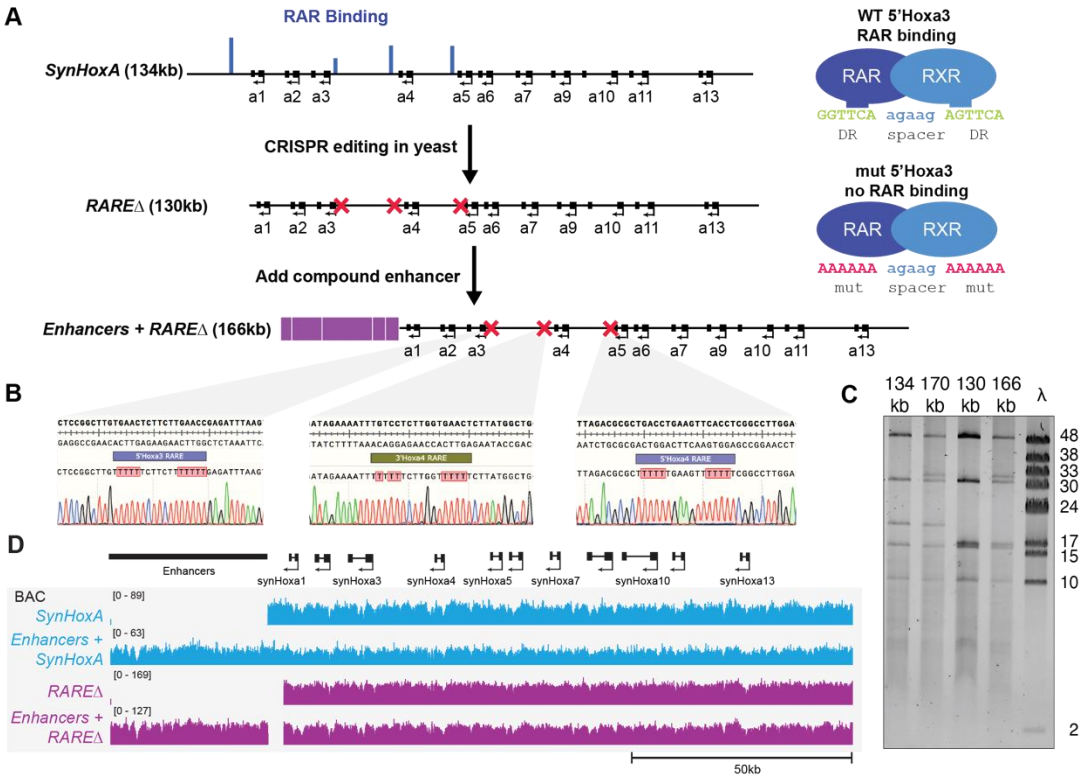


**Figure S19: Build of *RARE*Δ *SynHoxA* assemblons.**

(A) Schematic of assembly strategy for 130kb *RARE*Δ *SynHoxA* and 166kb *Enhancers+RARE*Δ *SynHoxA*. Nature of the RARE mutations is shown on the right. RAR binding data comes from previously published reports. (see Methods) (B) Sanger sequencing traces confirmed precise CRISPR editing of RAREs in yeast. (C) *SynHoxA* assemblon BACs purified from *E.coli* were digested with *PvuI* and separated using FIGE. Lambda monocut ladder sizes are indicated in kb. Bands correspond to expected fragment lengths. (D) Sequencing data of assemblon BACs purified from *E. coli* aligned to a custom mm10 reference genome. Positions of the enhancers and protein coding genes are shown in black.

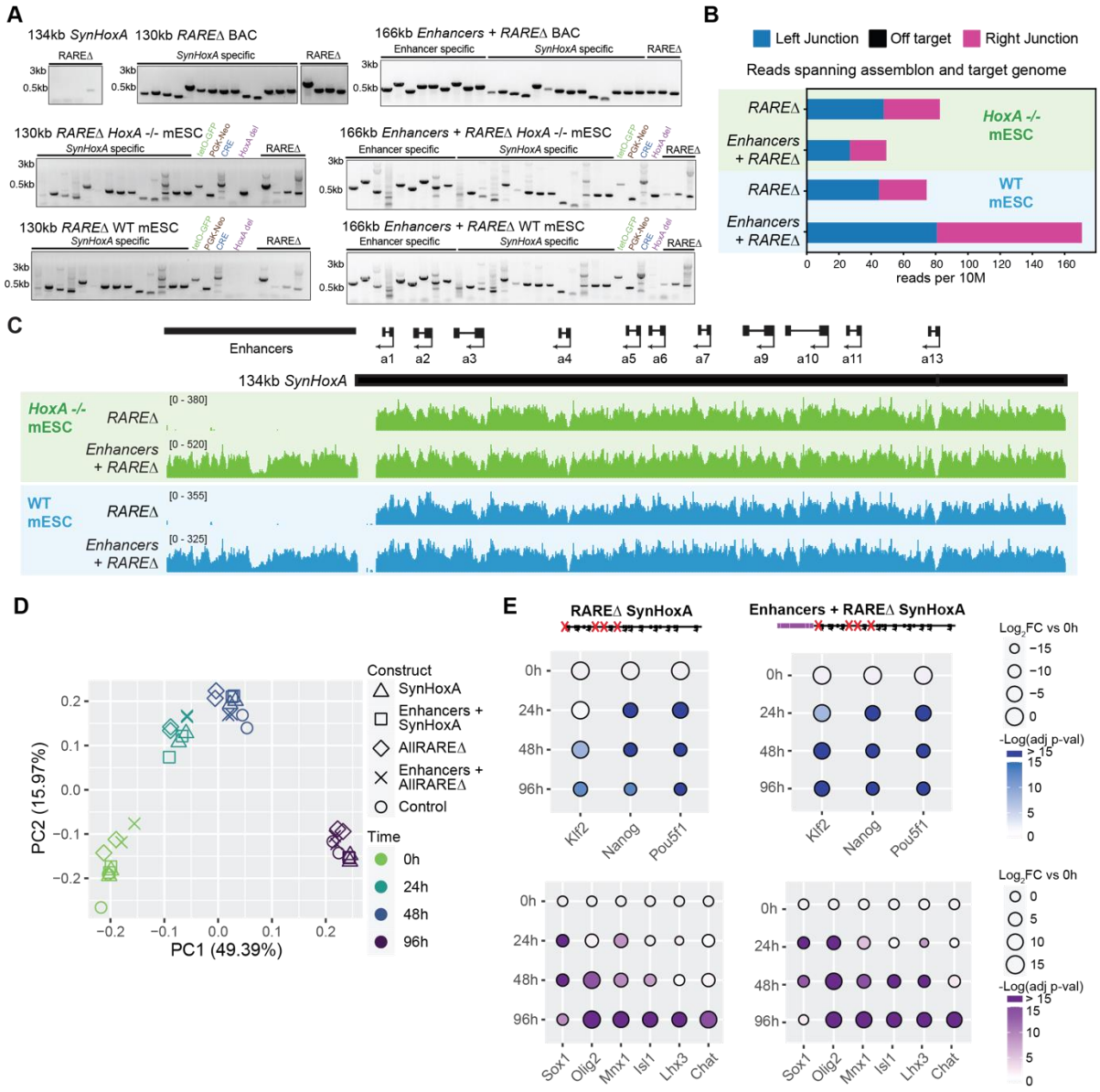


**Figure S20: Delivery of *RARE*Δ assemblons to WT and *HoxA*$^{-/-}$ mESCs.**
(A) Genotyping PCRs separated on an agarose gel. RAREΔ specific primers are used to verify presence of the RARE mutations. Clones were also screened using *SynHoxA* specific primers that span the length of the assembly, primers specific to enhancer junctions, primers that span novel junctions formed with the genome (tetO-GFP, PGK-Neo) and primers that confirm overwriting of the *Cre* gene. In addition, the presence or absence of the endogenous *HoxA* cluster deletion was confirmed using deletion specific primers. (B) Only the expected junctions spanning the synthetic assemblon and the host genome were observed in next generation sequencing data with no off-target integrations. (C) Positions of the enhancers and protein coding

genes are shown in black. Capture sequencing data is shown from WT and *HoxA*$^{-/-}$ mESC clones arising from delivery of *RARE*Δ *SynHoxA* assemblons. Sequencing data shown here are aligned to a custom reference genome (see Methods). (D) Principal Component Analysis (PCA) of the RNA-seq datasets reveals clustering largely by time during the differentiation protocol (each data point represents independent differentiations). This is true regardless of ectopic *SynHoxA* variants that are integrated. (E) The log2 fold change of pluripotency markers and MN differentiation genes from RNA-seq data (n=2). Pluripotency markers were downregulated and MN markers were upregulated during differentiation as expected for mESCs bearing *RARE*Δ *SynHoxA and Enhancers+RARE*Δ *SynHoxA.*

**References and Notes**


1. D. Duboule, The rise and fall of Hox gene clusters. *Development* **134**, 2549–2560 (2007). doi:10.1242/dev.001065 Medline

2. M. Kmita, D. Duboule, Organizing axes in time and space; 25 years of colinear tinkering. *Science* **301**, 331–333 (2003). doi:10.1126/science.1085753 Medline

3. W. McGinnis, R. Krumlauf, Homeobox genes and axial patterning. *Cell* **68**, 283–302 (1992). doi:10.1016/0092-8674(92)90471-N Medline

4. D. Duboule, G. Morata, Colinearity and functional hierarchy among genes of the homeotic complexes. *Trends Genet.* **10**, 358–364 (1994). doi:10.1016/0168-9525(94)90132-5 Medline

5. E. B. Lewis, A gene complex controlling segmentation in Drosophila. *Nature* **276**, 565–570 (1978). doi:10.1038/276565a0 Medline

6. N. Shah, S. Sukumar, The Hox genes and their roles in oncogenesis. *Nat. Rev. Cancer* **10**, 361–371 (2010). doi:10.1038/nrc2826 Medline

7. R. Margueron, D. Reinberg, The Polycomb complex PRC2 and its mark in life. *Nature* **469**, 343–349 (2011). doi:10.1038/nature09784 Medline

8. S. Mahony, E. O. Mazzoni, S. McCuine, R. A. Young, H. Wichterle, D. K. Gifford, Ligand-dependent dynamics of retinoic acid receptor binding during early neurogenesis. *Genome Biol.* **12**, R2 (2011). doi:10.1186/gb-2011-12-1-r2 Medline

9. C. Nolte, B. De Kumar, R. Krumlauf, Hox genes: Downstream “effectors” of retinoic acid signaling in vertebrate embryogenesis. *Genesis* **57**, e23306 (2019). doi:10.1002/dvg.23306 Medline

10. E. O. Mazzoni, S. Mahony, M. Peljto, T. Patel, S. R. Thornton, S. McCuine, C. Reeder, L. A. Boyer, R. A. Young, D. K. Gifford, H. Wichterle, Saltatory remodeling of Hox chromatin in response to rostrocaudal patterning signals. *Nat. Neurosci.* **16**, 1191–1198 (2013). doi:10.1038/nn.3490 Medline

11. D. Noordermeer, M. Leleu, E. Splinter, J. Rougemont, W. De Laat, D. Duboule, The dynamic architecture of Hox gene clusters. *Science* **334**, 222–225 (2011). doi:10.1126/science.1207194 Medline

12. N. Soshnikova, D. Duboule, Epigenetic temporal control of mouse Hox genes in vivo. *Science* **324**, 1320–1323 (2009). doi:10.1126/science.1171468 Medline

13. V. Dupé, M. Davenne, J. Brocard, P. Dollé, M. Mark, A. Dierich, P. Chambon, F. M. Rijli, In vivo functional analysis of the Hoxa-1 3′ retinoic acid response element (3'RARE). *Development* **124**, 399–410 (1997). doi:10.1242/dev.124.2.399 Medline

14. M. Frasch, X. Chen, T. Lufkin, Evolutionary-conserved enhancers direct region-specific expression of the murine Hoxa-1 and Hoxa-2 loci in both mice and Drosophila. *Development* **121**, 957–974 (1995). doi:10.1242/dev.121.4.957 Medline

15. N. Lonfat, T. Montavon, F. Darbellay, S. Gitto, D. Duboule, Convergent evolution of complex regulatory landscapes and pleiotropy at Hox loci. *Science* **346**, 1004–1006 (2014). doi:10.1126/science.1257493 Medline

16. T. Montavon, D. Duboule, Chromatin organization and global regulation of Hox gene clusters. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20120367 (2013). doi:10.1098/rstb.2012.0367 Medline

17. T. Montavon, N. Soshnikova, Hox gene regulation and timing in embryogenesis. *Semin. Cell Dev. Biol.* **34**, 76–84 (2014). doi:10.1016/j.semcdb.2014.06.005 Medline

18. T. Montavon, N. Soshnikova, B. Mascrez, E. Joye, L. Thevenet, E. Splinter, W. de Laat, F. Spitz, D. Duboule, A regulatory archipelago controls Hox genes transcription in digits. *Cell* **147**, 1132–1145 (2011). doi:10.1016/j.cell.2011.10.023 Medline

19. S. Berlivet, D. Paquette, A. Dumouchel, D. Langlais, J. Dostie, M. Kmita, Clustering of tissue-specific sub-TADs accompanies the regulation of HoxA genes in developing limbs. *PLOS Genet.* **9**, e1004018 (2013). doi:10.1371/journal.pgen.1004018 Medline

20. K. Cao, C. K. Collings, S. A. Marshall, M. A. Morgan, E. J. Rendleman, L. Wang, C. C. Sze, T. Sun, E. T. Bartom, A. Shilatifard, SET1A/COMPASS and shadow enhancers in the regulation of homeotic gene expression. *Genes Dev.* **31**, 787–801 (2017). doi:10.1101/gad.294744.116 Medline

21. B. De Kumar, M. E. Parrish, B. D. Slaughter, J. R. Unruh, M. Gogol, C. Seidel, A. Paulson, H. Li, K. Gaudenz, A. Peak, W. McDowell, B. Fleharty, Y. Ahn, C. Lin, E. Smith, A. Shilatifard, R. Krumlauf, Analysis of dynamic changes in retinoid-induced transcription and epigenetic profiles of murine Hox clusters in ES cells. *Genome Res.* **25**, 1229–1243 (2015). doi:10.1101/gr.184978.114 Medline

22. R. Neijts, S. Amin, C. van Rooijen, S. Tan, M. P. Creyghton, W. de Laat, J. Deschamps, Polarized regulatory landscape and Wnt responsiveness underlie Hox activation in embryos. *Genes Dev.* **30**, 1937–1942 (2016). doi:10.1101/gad.285767.116 Medline

23. V. Narendra, P. P. Rocha, D. An, R. Raviram, J. A. Skok, E. O. Mazzoni, D. Reinberg, CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* **347**, 1017–1021 (2015). doi:10.1126/science.1262088 Medline

24. V. Narendra, M. Bulajić, J. Dekker, E. O. Mazzoni, D. Reinberg, CTCF-mediated topological boundaries during development foster appropriate gene regulation. *Genes Dev.* **30**, 2657–2662 (2016). doi:10.1101/gad.288324.116 Medline

25. N. Ostrov, J. Beal, T. Ellis, D. B. Gordon, B. J. Karas, H. H. Lee, S. C. Lenaghan, J. A. Schloss, G. Stracquadanio, A. Trefzer, J. S. Bader, G. M. Church, C. M. Coelho, J. W. Efcavitch, M. Güell, L. A. Mitchell, A. A. K. Nielsen, B. Peck, A. C. Smith, C. N. Stewart Jr., H. Tekotte, Technological challenges and milestones for writing genomes. *Science* **366**, 310–312 (2019). doi:10.1126/science.aay0339 Medline

26. M. Gasperini, L. Starita, J. Shendure, The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.* **11**, 1782–1787 (2016). doi:10.1038/nprot.2016.135 Medline

27. J. A. Lehoczky, J. W. Innis, BAC transgenic analysis reveals enhancers sufficient for Hoxa13 and neighborhood gene expression in mouse embryonic distal limbs and genital bud. *Evol. Dev.* **10**, 421–432 (2008). doi:10.1111/j.1525-142X.2008.00253.x Medline

28. F. Spitz, F. Gonzalez, D. Duboule, A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* **113**, 405–417 (2003). doi:10.1016/S0092-8674(03)00310-6 Medline

29. F. Spitz, F. Gonzalez, C. Peichel, T. F. Vogt, D. Duboule, J. Zákány, Large scale transgenic and cluster deletion analysis of the HoxD complex separate an ancestral regulatory module from evolutionary innovations. *Genes Dev.* **15**, 2209–2214 (2001). doi:10.1101/gad.205701 Medline

30. K. R. Peterson, Q. L. Li, C. H. Clegg, T. Furukawa, P. A. Navas, E. J. Norton, T. G. Kimbrough, G. Stamatoyannopoulos, Use of yeast artificial chromosomes (YACs) in studies of mammalian development: Production of beta-globin locus YAC mice carrying human globin developmental mutants. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 5655–5659 (1995). doi:10.1073/pnas.92.12.5655 Medline

31. N. Heintz, BAC to the future: The use of bac transgenic mice for neuroscience research. *Nat. Rev. Neurosci.* **2**, 861–870 (2001). doi:10.1038/35104049 Medline

32. F. G. Liberante, T. Ellis, From kilobases to megabases: Design and delivery of large DNA constructs into mammalian genomes. *Curr. Opin. Syst. Biol.* **25**, 1–10 (2021). doi:10.1016/j.coisb.2020.11.003

33. H. A. Wallace, F. Marques-Kranc, M. Richardson, F. Luna-Crespo, J. A. Sharpe, J. Hughes, W. G. Wood, D. R. Higgs, A. J. H. Smith, Manipulating the mouse genome to engineer precise functional syntenic replacements with human sequence. *Cell* **128**, 197–209 (2007). doi:10.1016/j.cell.2006.11.044 Medline

34. N. S. McCarty, A. E. Graham, L. Studená, R. Ledesma-Amaro, Multiplexed CRISPR technologies for gene editing and transcriptional regulation. *Nat. Commun.* **11**, 1281 (2020). doi:10.1038/s41467-020-15053-x Medline

35. K. Kraft, S. Geuer, A. J. Will, W. L. Chan, C. Paliou, M. Borschiwer, I. Harabula, L. Wittler, M. Franke, D. M. Ibrahim, B. K. Kragesteen, M. Spielmann, S. Mundlos, D. G. Lupiáñez, G. Andrey, Deletions, Inversions, Duplications: Engineering of Structural Variants using CRISPR/Cas in Mice. *Cell Rep.* **10**, 833–839 (2015). doi:10.1016/j.celrep.2015.01.016 Medline

36. L. A. Mitchell, L. H. McCulloch, S. Pinglay, H. Berger, N. Bosco, R. Brosh, M. Bulajić, E. Huang, M. S. Hogan, J. A. Martin, E. O. Mazzoni, T. Davoli, M. T. Maurano, J. D.

Boeke, De novo assembly and delivery to mouse cells of a 101 kb functional human gene. *Genetics* **218**, iyab038 (2021). doi:10.1093/genetics/iyab038 Medline

37. J. E. DiCarlo, J. E. Norville, P. Mali, X. Rios, J. Aach, G. M. Church, Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Res.* **41**, 4336–4343 (2013). doi:10.1093/nar/gkt135 Medline

38. M. Iacovino, D. Bosnakovski, H. Fey, D. Rux, G. Bajwa, E. Mahen, A. Mitanoska, Z. Xu, M. Kyba, Inducible cassette exchange: A rapid and efficient system enabling conditional gene expression in embryonic stem and primary cells. *Stem Cells* **29**, 1580–1588 (2011). doi:10.1002/stem.715 Medline

39. M. Jasin, M. E. Moynahan, C. Richardson, Targeted transgenesis. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 8804–8808 (1996). doi:10.1073/pnas.93.17.8804 Medline

40. M. Gasperini, G. M. Findlay, A. McKenna, J. H. Milbank, C. Lee, M. D. Zhang, D. A. Cusanovich, J. Shendure, CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *Am. J. Hum. Genet.* **101**, 192–205 (2017). doi:10.1016/j.ajhg.2017.06.010 Medline

41. R. Brosh, J. M. Laurent, R. Ordoñez, E. Huang, M. S. Hogan, A. M. Hitchcock, L. A. Mitchell, S. Pinglay, J. A. Cadley, R. D. Luther, D. M. Truong, J. D. Boeke, M. T. Maurano, A versatile platform for locus-scale genome rewriting and verification. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2023952118 (2021). doi:10.1073/pnas.2023952118 Medline

42. H. Wichterle, I. Lieberam, J. A. Porter, T. M. Jessell, Directed differentiation of embryonic stem cells into motor neurons. *Cell* **110**, 385–397 (2002). doi:10.1016/S0092-8674(02)00835-8 Medline

43. M. Peljto, H. Wichterle, Programming embryonic stem cells to neuronal subtypes. *Curr. Opin. Neurobiol.* **21**, 43–51 (2011). doi:10.1016/j.conb.2010.09.012 Medline

44. B. Aydin, A. Kakumanu, M. Rossillo, M. Moreno-Estellés, G. Garipler, N. Ringstad, N. Flames, S. Mahony, E. O. Mazzoni, Proneural factors Ascl1 and Neurog2 contribute to neuronal subtype identities by establishing distinct chromatin landscapes. *Nat. Neurosci.* **22**, 897–908 (2019). doi:10.1038/s41593-019-0399-y Medline

45. M. Bulajić, D. Srivastava, J. S. Dasen, H. Wichterle, S. Mahony, E. O. Mazzoni, Differential abilities to engage inaccessible chromatin diversify vertebrate Hox binding patterns. *Development* **147**, dev.194761 (2020). doi:10.1242/dev.194761 Medline

46. G. Su, W. Wang, J. Chen, M. Liu, J. Zheng, D. Guo, J. Bi, Z. Zhao, J. Shi, L. Zhang, W. Lu, CTCF-binding element regulates ESC differentiation via orchestrating long-range chromatin interaction between enhancers and HoxA. *J. Biol. Chem.* **296**, 100413 (2021). doi:10.1016/j.jbc.2021.100413 Medline

47. R. Freitas, C. Gómez-Marín, J. M. Wilson, F. Casares, J. L. Gómez-Skarmeta, Hoxd13 contribution to the evolution of vertebrate appendages. *Dev. Cell* **23**, 1219–1229 (2012). doi:10.1016/j.devcel.2012.10.015 Medline

48. H. Sato, S. Das, R. H. Singer, M. Vera, Imaging of DNA and RNA in Living Eukaryotic Cells to Reveal Spatiotemporal Dynamics of Gene Expression. *Annu. Rev. Biochem.* **89**, 159–187 (2020). doi:10.1146/annurev-biochem-011520-104955 Medline

49. L. Beccari, N. Moris, M. Girgin, D. A. Turner, P. Baillie-Johnson, A.-C. Cossy, M. P. Lutolf, D. Duboule, A. M. Arias, Multi-axial self-organization properties of mouse embryonic stem cells into gastruloids. *Nature* **562**, 272–276 (2018). doi:10.1038/s41586-018-0578-0 Medline

50. K. A. Ganzinger, P. Schwille, More from less - bottom-up reconstitution of cell biology. *J. Cell Sci.* **132**, jcs227488 (2019). doi:10.1242/jcs.227488 Medline

51. A. P. Liu, D. A. Fletcher, Biology under construction: In vitro reconstitution of cellular function. *Nat. Rev. Mol. Cell Biol.* **10**, 644–650 (2009). doi:10.1038/nrm2746 Medline

52. R. D. Gietz, R. H. Schiestl, High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31–34 (2007). doi:10.1038/nprot.2007.13 Medline

53. J. S. Dymond, S. M. Richardson, C. E. Coombes, T. Babatz, H. Muller, N. Annaluru, W. J. Blake, J. W. Schwerzmann, J. Dai, D. L. Lindstrom, A. C. Boeke, D. E. Gottschling, S. Chandrasegaran, J. S. Bader, J. D. Boeke, Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature* **477**, 471–476 (2011). doi:10.1038/nature10403 Medline

54. N. Agmon, J. Temple, Z. Tang, T. Schraink, M. Baron, J. Chen, P. Mita, J. A. Martin, B. P. Tu, I. Yanai, D. Fenyö, J. D. Boeke, Phylogenetic debugging of a complete human biosynthetic pathway transplanted into yeast. *Nucleic Acids Res.* **48**, 486–499 (2020). Medline

55. J. Sambrook, D. W. Russell, Isolation of BAC DNA from Small-scale Cultures. *CSH Protoc.* **2006**, XXX (2006).

56. A. Untergasser, I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm, S. G. Rozen, Primer3—New capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012). doi:10.1093/nar/gks596 Medline

57. J. P. Concordet, M. Haeussler, CRISPOR: Intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.* **46**, W242–W245 (2018). doi:10.1093/nar/gky354 Medline

58. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014). doi:10.1093/bioinformatics/btu170 Medline

59. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). doi:10.1093/bioinformatics/btp324 Medline

60. G. G. Faust, I. M. Hall, SAMBLASTER: Fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014). doi:10.1093/bioinformatics/btu314 Medline

61. S. Neph, M. S. Kuehn, A. P. Reynolds, E. Haugen, R. E. Thurman, A. K. Johnson, E. Rynes, M. T. Maurano, J. Vierstra, S. Thomas, R. Sandstrom, R. Humbert, J. A. Stamatoyannopoulos, BEDOPS: High-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012). doi:10.1093/bioinformatics/bts277 Medline

62. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). doi:10.1038/nmeth.1923 Medline

63. F. Ramírez, F. Dündar, S. Diehl, B. A. Grüning, T. Manke, deepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014). doi:10.1093/nar/gku365 Medline

64. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi:10.1093/bioinformatics/btp352 Medline

65. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011). doi:10.1038/nbt.1754 Medline

66. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). doi:10.1093/bioinformatics/btq033 Medline

67. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019). doi:10.1038/s41587-019-0201-4 Medline

68. Y. Liao, G. K. Smyth, W. Shi, The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **47**, e47 (2019). doi:10.1093/nar/gkz114 Medline

69. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014). doi:10.1186/s13059-014-0550-8 Medline

70. H. Wickham, *Use R!* (Springer, 2016).

71. Y. Zhang, G. Parmigiani, W. E. Johnson, *ComBat-seq*: Batch effect adjustment for RNA-seq count data. *NAR Genom. Bioinform.* **2**, lqaa078 (2020). doi:10.1093/nargab/lqaa078 Medline

72. Y. Zheng, F. Ay, S. Keles, Generative modeling of multi-mapping reads with mHi-C advances analysis of Hi-C studies. *eLife* **8**, e38070 (2019). doi:10.7554/eLife.38070 Medline

73. C. Lévy-Leduc, M. Delattre, T. Mary-Huard, S. Robin, Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics* **30**, i386–i392 (2014). doi:10.1093/bioinformatics/btu443 Medline

74. J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, B. Ren, Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012). doi:10.1038/nature11082 Medline

75. H. M. T. Choi, M. Schwarzkopf, M. E. Fornace, A. Acharya, G. Artavanis, J. Stegmaier, A. Cunha, N. A. Pierce, Third-generation *in situ* hybridization chain reaction: Multiplexed, quantitative, sensitive, versatile, robust. *Development* **145**, dev165753 (2018). doi:10.1242/dev.165753 Medline

76. C. Xia, H. P. Babcock, J. R. Moffitt, X. Zhuang, Multiplexed detection of RNA using MERFISH and branched DNA amplification. *Sci. Rep.* **9**, 7721 (2019). doi:10.1038/s41598-019-43943-8 Medline

77. C. McQuin, A. Goodman, V. Chernyshev, L. Kamentsky, B. A. Cimini, K. W. Karhohs, M. Doan, L. Ding, S. M. Rafelski, D. Thirstrup, W. Wiegraebe, S. Singh, T. Becker, J. C. Caicedo, A. E. Carpenter, CellProfiler 3.0: Next-generation image processing for biology. *PLOS Biol.* **16**, e2005970 (2018). doi:10.1371/journal.pbio.2005970 Medline

78. T. Lionnet, K. Czaplinski, X. Darzacq, Y. Shav-Tal, A. L. Wells, J. A. Chao, H. Y. Park, V. de Turris, M. Lopez-Jones, R. H. Singer, A transgenic mouse for in vivo detection of endogenous labeled mRNA. *Nat. Methods* **8**, 165–170 (2011). doi:10.1038/nmeth.1551 Medline

79. J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, A. Cardona, Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012). doi:10.1038/nmeth.2019 Medline