Supplementary Information for
Spatial reconstruction of single-cell gene expression

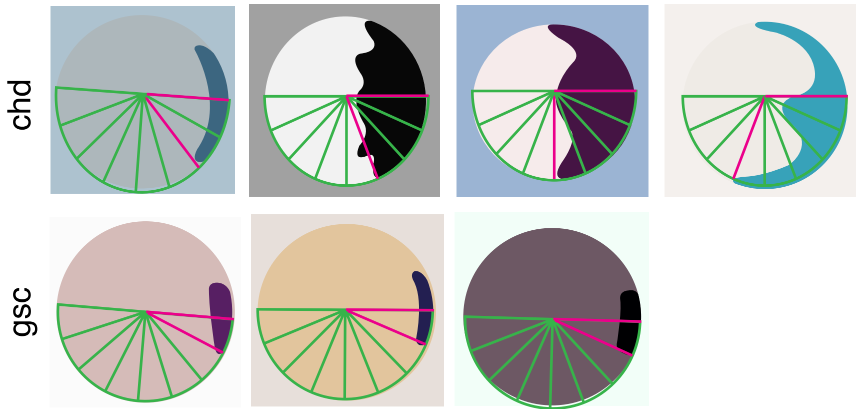
Rahul Satija and Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev

Supplementary Text: Spatially diverse landmark genes improve Seurat's mapping

To assess Seurat's sensitivity to the number and type of landmark genes composing our spatial reference map, we downsampled the number of landmark genes used as input to Seurat and performed a spatial power analysis. The achievable resolution of spatial mapping (*i.e.*, the number of bins that can be reliably defined) will depend on the number of spatially unique combinations of landmark gene expression. Thus, we reasoned that a smaller number of landmark genes spanning a diverse range of expression patterns would outperform a larger number of landmark genes with overlapping or redundant patterns. For each downsampling, we first selected random sets of 2, 4, 6, or all 9 archetypal groups, then sampled 2–45 landmark genes evenly across the selected archetypes, and used them to construct a reduced spatial reference map. We then repeated Seurat's mapping on each reduced spatial reference map and compared both the resulting cell mappings (as Euclidean centroid distance) and confidences (as overall shift in posterior probability) to those obtained using the full reference map (generated with 46 of the 47 landmark genes that met our variability requirements for inclusion in the archetypal clustering).

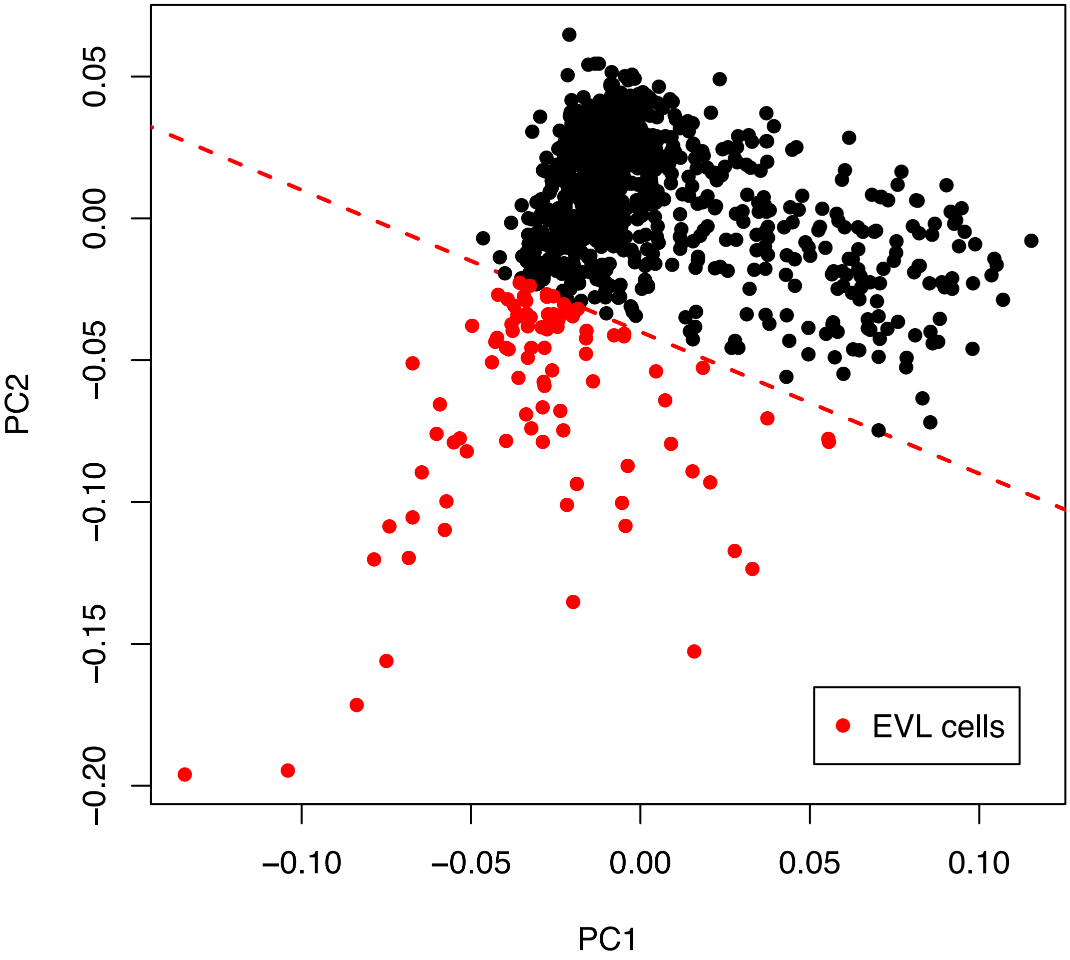
We obtained similar mappings from a reduced number of landmark genes, with quality affected by their number and spatial diversity. For example, even with reference maps from certain subsets of only 16 landmark genes, Seurat's cell mappings shifted by less than one bin on average, with this distance dropping to approximately half a bin and beginning to saturate after including 29 landmark genes (**Supplementary Fig. 7a**). However, this result held only if our genes were evenly sampled across all nine archetypes: we observed larger shifts when the reference map was constructed from the same number of landmark genes, but sampled from a smaller number of archetypes. Thus, spatial diversity in our landmarks provides the greatest benefit to Seurat's mapping. Nevertheless, increasing the number of landmarks, even if they have overlapping expression patterns, does continue to improve Seurat's confidence in the resulting mappings (**Supplementary Fig. 7b**). To further investigate the effect of redundant landmarks, we considered the contribution of 2 sets of 4 landmark genes with identical expression patterns, in a reference map that we collapsed to the three bins that were defined by these landmarks. We found that 2X redundancy (*i.e.* two genes with overlapping expression) improves Seurat's mapping, but increased redundancy led to diminishing returns (**Supplementary Fig. 7c**).

Supplementary Figure 1: Variability in published *in situ* images.



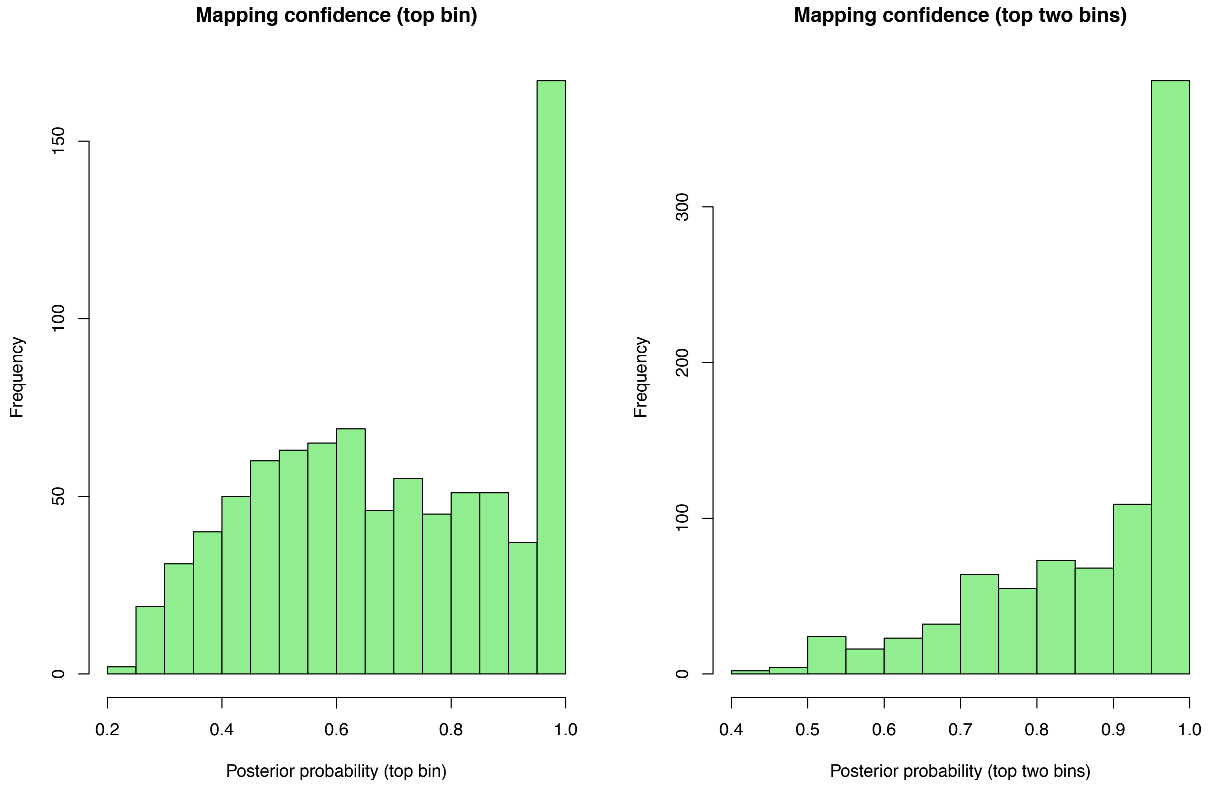
Cartoon representations of four published *in situ* images for *chd* (Gerdes et al., 2007 Fig. 5c; Sidi et al., 2003 Fig. 2D; Gilardelli et al., 2004 Fig. 5D; Maegawa et al., 2006 Fig. 5B) and cartoon representations of three published *in situ* images for *gsc* (Tian et. al, 2008 Fig. 1B; Thisse & Thisse, 2004; Du et al., 2012 Fig. 9A), identified as 50% epiboly in their publications and part of our landmark data set (references in **Supplementary Table 1**). Scoring grids used to determine the spatial extent of gene expression are superimposed on top in green, with pink lines demarcating our interpretation of the extent of gene expression around the circumference as determined from the staining. For many genes, such as *gsc*, patterns are highly reproduced between studies—for all three images we would interpret as staining 1 bin. However, other genes, such as *chd*, show high variability, likely due to differences in embryo staging, extent of stain development, and imaging conditions. For *chd*, from the cited images, we would score its expression as occupying 2, 3, 4, or 5 bins (from left to right).

Supplementary Figure 2: Removal of EVL cells.



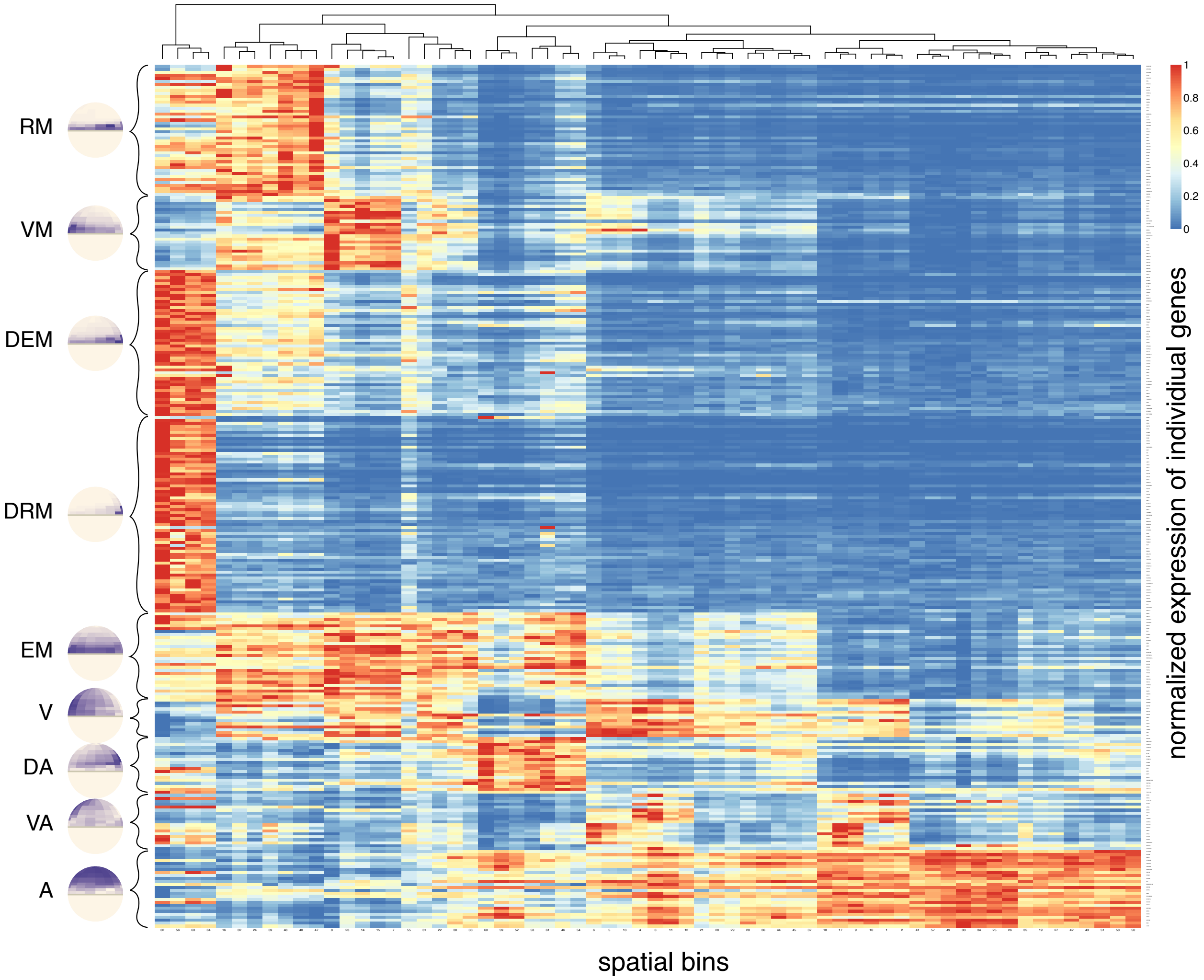
A single layer of enveloping layer (EVL) cells ubiquitously covers the outside of the embryo, but these cells do not express the same group of genes used to construct our spatial map. Thus, we exclude them from the study. The cells were identified primarily based on their strong loadings for the second principal component, which was defined by canonical markers of the EVL (e.g. *krt18, krt4, cldne*). We chose to select a restrictive cutoff to minimize the potential contribution of EVL cells to our dataset.

Supplementary Figure 3: Mapping confidence.



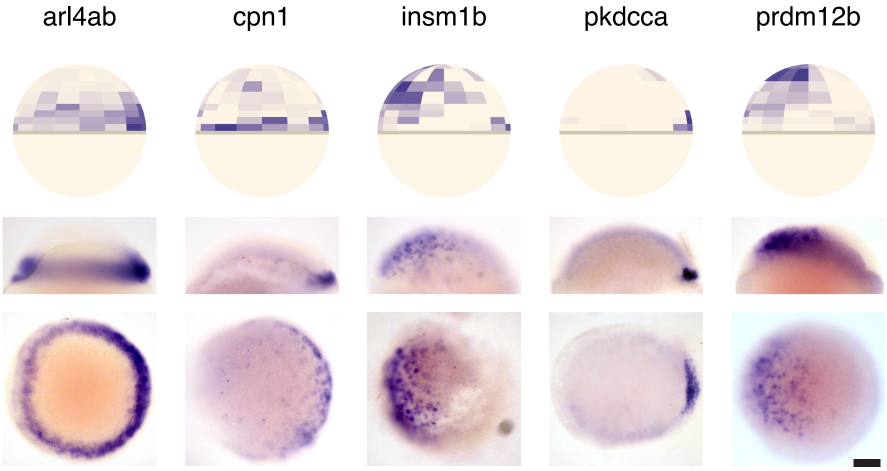
Seurat probabilistically assigns cells to one or more bins by determining a posterior probability that a cell originated from each of 64 bins in our model. As a measure of how confidently cells were mapped, histograms of the posterior probability of the most likely bin (**left**) or the sum of the two most likely bins (**right**) for each cell are displayed. Seurat mapped the majority of cells to one or two bins with high confidence (p>0.9), (24% for a single bin, 59% for two bins, which are typically adjacent).

Supplementary Figure 4: Spatial prediction clustering and archetype prediction.



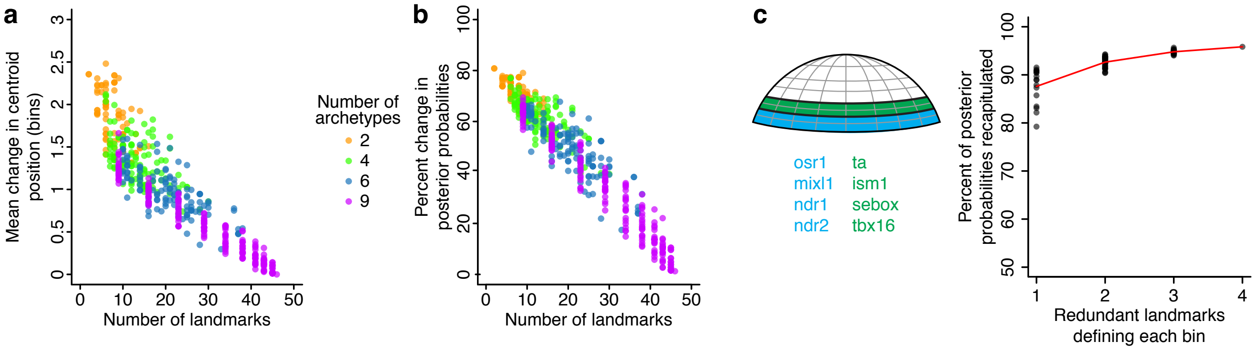
The predicted spatial pattern was determined from the imputed expression values for the 290 spatially variable genes in the data set (**Methods**). The clustered matrix has genes as rows, and each of the 64 bins as columns. We used k-means clustering to cluster genes into 9 clusters of 'archetypal' expression patterns, using hierarchical clustering to group the columns for visualization. The averaged expression pattern of all genes in each archetype are displayed to the left. The archetypes are: restricted margin (RM), ventral margin (VM), dorsally enriched margin (DEM), dorsally restricted margin (DRM), extended margin (EM), ventral (V), dorsal animal (DA), ventral animal (VA), animal (A).

Supplementary Figure 5: Additional *in situ* patterns for Figure 4.



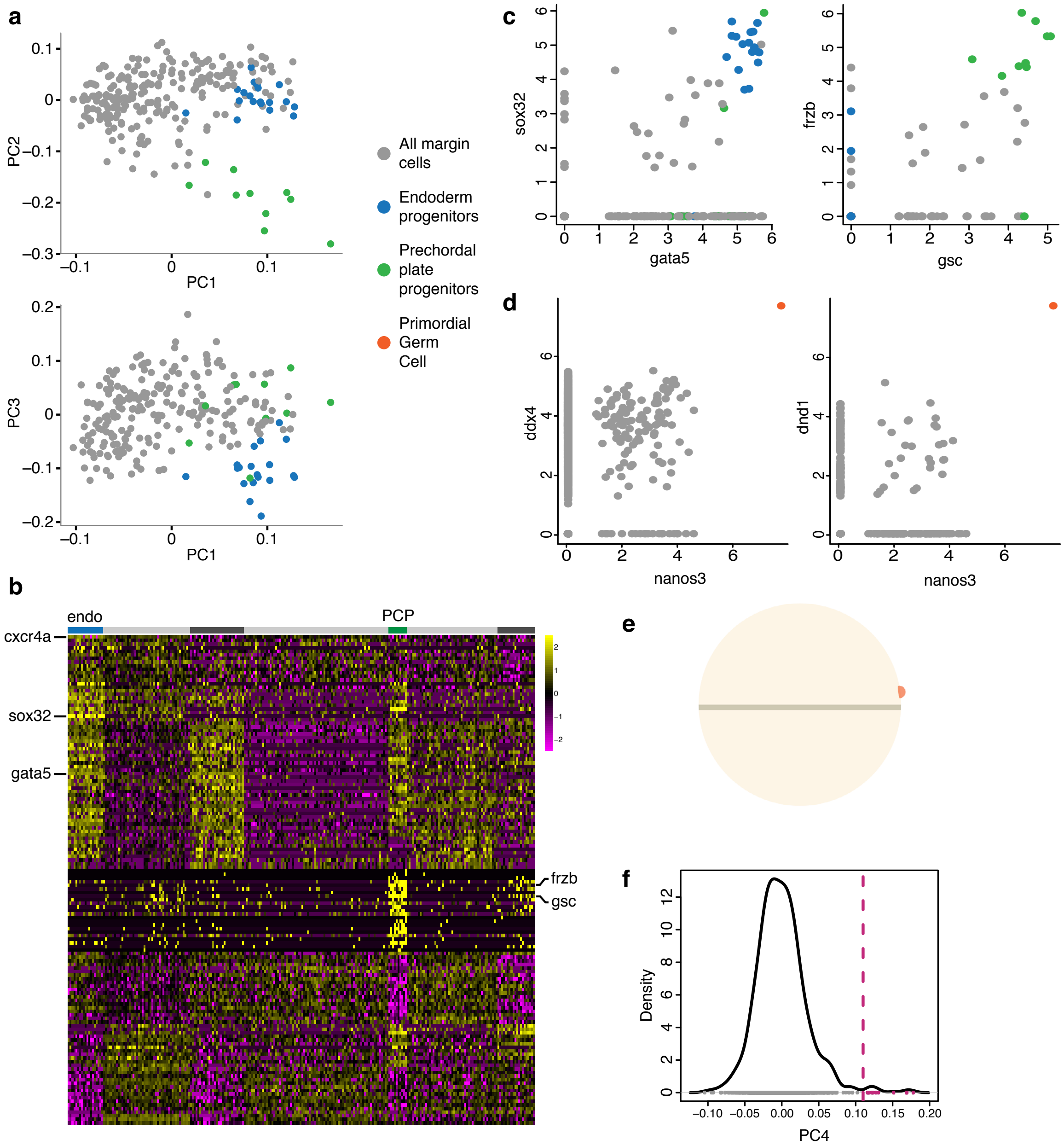
Five additional *in situs* for which there was no published 50% epiboly pattern (assessed Sep 4, 2014). Top to bottom: Seurat's predicted expression pattern, lateral view (dorsal to the right), animal cap view (dorsal to the right). Scale bar represents 100 µm.

Supplementary Figure 6: Spatially diverse landmark genes improve Seurat's mapping power.



(**a, b**) Spatial mapping power analysis. Shown are the mean change in centroid positions (**Y axis, a**) and the percentage change in posterior probabilities (**Y axis, b**) of cells mapped by Seurat when using a reference map from 500 limited subsets of landmark genes. To generate the limited subsets, first, 2 (**orange**), 4 (**green**), 6 (**blue**), or 9 (**purple**) archetypes were randomly chosen, and then 2–45 landmark genes (**X axis**) were randomly chosen evenly across those archetypes. The cells were then remapped with the reference map based on the landmark subset. (c) Contribution of redundant landmark patterns. We generated a reduced resolution reference map, based on partitioning the embryo into only 3 bins (**left: white, blue, and green**). We then selected sets of landmark genes with identical expression patterns that defined these bins—4 landmark genes that had the expression pattern of the blue bin, and 4 that had the expression pattern of the blue + green bins. We then mapped cells within these 3 bins using a reference map with the full set of landmarks, or with every possible combination of 1, 2, 3, or 4 of the landmarks from each of the two redundant expression patterns. Shown are the percent of posterior probabilities recovered (**Y axis**) at each number of markers per bin (**X axis**), as compared to the full set of landmarks. Inclusion of two landmark genes with redundant patterns provides greater accuracy than inclusion of only one, but additional redundancy provides only diminishing returns.

Supplementary Figure 7: Identification of rare subpopulations.



(**a**) Two unsupervised analyses (PCA analysis, k-means) both identify known subpopulations near the margin. The prechordal plate progenitors (green) and endodermal progenitors (blue) as identified by k-means clustering separate well based on principal components 2 and 3, respectively. (**b**) k-means clustering identification of the endodermal progenitors (blue) and prechordal plate progenitors (green). Classical markers of these populations were used to confirm the biologically identify of these clusters, and are displayed to the left and right of the heatmap. (**c**) Endodermal progenitors (blue) and prechordal plate progenitors (green) identified by k-means clustering would also have largely been identified via gating approaches based on classical markers of the endodermal progenitors (*sox32* and *gata5*, **left**) and prechordal plate progenitors (*gsc* and *frzb*, **right**). (**d**) Identification via a supervised gating approach of a primordial germ cell (PGC) based on its classical markers, *nanos3*, *ddx4/vasa*, and *dnd1*. (**e**) Seurat correctly maps the identified PGC near the embryonic margin. (**f**) Identification of the 'apoptotic-like' cell population (magenta) by thresholding along principal component 4.

Supplementary Table 1
Landmark genes and references to the images used in this study

Landmark Gene	**Published *in situ* images used to generate spatial map**
admp	(Bennett et al. 2007) Fig. S1, (Lele et al. 2001) Fig. 1
aplnrb	(Scott et al. 2007) Fig. 3B, (Pauli et al. 2014) Fig. S17A
axin2	(B. Thisse & C. Thisse 2004)
bambia	(Bennett et al. 2007) Fig. S1
bmp2b	(Okuda et al. 2010) Fig. 4b, (Yimlamai et al. 2005) Fig. 7D, (Gilardelli et al. 2004) Fig. 5J, (Sidi et al. 2003) Fig. 2A, (B. Thisse et al. 2001)
bmp4	(Zon et al. 1997) Fig. 6C, (Sidi et al. 2003) Figs. 2B, 2J, (B. Thisse & C. Thisse 2005)
bmp7a	(Okuda et al. 2010) Fig. 4e, (Jurynec & Grunwald 2010) Fig. 3C
cdx4	(B. Thisse et al. 2001), (Ramel et al. 2005) Fig. 4S, (Davidson et al. 2003) Fig. 2d, (Lin et al. 2007) Fig. 4G
chd	(Sidi et al. 2003) Fig. 2D, (Aamar & Dawid 2010) Fig. 3A, (Maegawa et al. 2006) Fig. 5B, (Wilm & Solnica-Krezel 2005) Fig. 4K, (Gilardelli et al. 2004) Fig. 5D, (Okuda et al. 2010) Fig. 4l, (Flores et al. 2008) Fig. 2d
cst3	(Bennett et al. 2007) Fig. S1
eve1	(Hammerschmidt et al. 1996) Fig. 2B, (Lyman Gingerich et al. 2005) Fig. 2e, (Fukazawa et al. 2010) Fig. 1K, (Fisher et al. 1997) Fig. 5E
fgf8a	(Koshida et al. 2002) Fig. 3C, (Walshe et al. 2002) Fig. 1B, (Fürthauer et al. 1997) Fig. 2B–C
foxb1a	(B. Thisse et al. 2001), (Jurynec & Grunwald 2010) Fig. 3I, 3K,
foxd3	(Wang et al. 2011) Fig. 3D, 3G
foxd5	(Bennett et al. 2007) Fig. S1
foxi1	(Dee et al. 2007) Fig. 1A, (Jia et al. 2012) Fig. 3A, (Xie et al. 2011) Fig. 6A5
gata2a	(Dee et al. 2007) Fig. 1A, (Mei et al. 2009) Fig. 5C
gata5	(Reiter et al. 2001) Fig. 2A, (Warga & Kane 2003) Fig. 5I, (Aoki et al. 2002) Fig. 1J
gsc	(Flores et al. 2008) Fig. 2e, (Tian et al. 2008) Fig. 1B, (Du et al. 2012) Fig. 9A, (B. Thisse & C. Thisse 2004), (Inbal et al. 2006) Fig. 1b
her1	(Bennett et al. 2007) Fig. S1, (B. Thisse & C. Thisse 2005)
hes6	(Bennett et al. 2007) Fig. S1, (Kawamura et al. 2005) Fig. 1A
id3	(Li et al. 2010) Fig. 1A
ism1	(Bennett et al. 2007) Fig. S1
lft1	Katherine Rogers, unpublished data
lft2	Katherine Rogers, unpublished data
lhx1a	(Toyama & Dawid 1997), Fig. 2A–B

mixl1 (mixer)	(Aoki et al. 2002) Fig. 1I, (Kunwar et al. 2003) Fig. 1C
ndr1 (sqt)	Katherine Rogers, unpublished data
ndr2 (cyc)	Katherine Rogers, unpublished data
nog1	(Dal-Pra et al. 2011) Fig. 2B
noto (flh)	Katherine Rogers, unpublished data, (Gilardelli et al. 2004) Fig. 5G, (Tian et al. 2003) Fig. 8A
osr1	(Mudumana et al. 2008) Fig. 1C
otx1a	(Bennett et al. 2007) Fig. S1
otx1b	(Bennett et al. 2007) Fig. S1
sebox	(Tseng et al. 2011) Fig. 4D-1
snai1a	(B. Thisse et al. 2001), (Zhao et al. 2003) Fig. 6A
sox3	(Kudoh et al. 2004) Fig. 1A
sp5l	(Zhao et al. 2003) Figs. 2D & H
szl	(Wilm & Solnica-Krezel 2005) Figs. 4M & 6E, (Dal-Pra et al. 2006) Fig. 1G, (Wu et al. 2012) Fig. 4Ba
ta (ntl)	(Bennett et al. 2007) Fig. S1, (Du et al. 2012) Fig. 9C”
tbx16 (spt)	(He et al. 2014) Figs. 5H & H’
tph1b	(Bennett et al. 2007) Fig. S1
ved	(B. Thisse et al. 2001), (Gilardelli et al. 2004) Fig. 2J, (Kapp et al. 2013) Fig. 4G
vent	(Gilardelli et al. 2004) Fig. 2L, (Kawahara et al. 2000) Fig. G, (Reim & Brand 2006) Fig. 3E, (Krens et al. 2008) Fig. 10J
vox	(Gilardelli et al. 2004) Fig. 2K, (Reim & Brand 2006) Fig. 3D, (Krens et al. 2008) Fig. 10G, (Shimizu et al. 2005) Fig. 2Ba, (Kapp et al. 2013) Fig. 4E, (He et al. 2014) Fig. 6J
wnt8a	(Wilm & Solnica-Krezel 2005) Fig. 5C, (Xie et al. 2011) Fig. 6B1, (Yao et al. 2010) Fig. 5D
wnt11	(Seo et al. 2010) Fig. 6A

Aamar, E. & Dawid, I.B., 2010. *sox17* and *chordin* are required for formation of Kupffer's vesicle and left-right asymmetry determination in zebrafish. *Developmental Dynamics*, 239(11), pp. 2980–2988.

Aoki, T.O. et al., 2002. Molecular integration of casanova in the Nodal signalling pathway controlling endoderm formation. *Development*, 129(2), pp. 275–286.

Bennett, J.T. et al., 2007. Nodal signaling activates differentiation genes during zebrafish gastrulation. *Developmental biology*, 304(2), pp. 525–540.

Dal-Pra, S. et al., 2006. Noggin1 and Follistatin-like2 function redundantly to Chordin to antagonize BMP activity. *Developmental biology*, 298(2), pp. 514–526.

Dal-Pra, S., Thisse, C. & Thisse, B., 2011. FoxA transcription factors are essential for the development of dorsal axial structures. *Developmental biology*, 350(2), pp. 484–495.

Davidson, A.J. et al., 2003. *cdx4* mutants fail to specify blood progenitors and can be rescued by multiple *hox* genes. *Nature*, 425(6955), pp. 300–306.

Dee, C.T. et al., 2007. A change in response to Bmp signalling precedes ectodermal fate choice. *The International journal of developmental biology*, 51(1), pp. 79–84.

Du, S. et al., 2012. Differential regulation of epiboly initiation and progression by zebrafish Eomesodermin A. *Developmental biology*, 362(1), pp. 11–23.

Fisher, S., Amacher, S.L. & Halpern, M.E., 1997. Loss of *cerebum* function ventralizes the zebrafish embryo. *Development*, 124(7), pp. 1301–1311.

Flores, M.V.C. et al., 2008. Osteogenic transcription factor Runx2 is a maternal determinant of dorsoventral patterning in zebrafish. *Nature cell biology*, 10(3), pp. 346–352.

Fukazawa, C. et al., 2010. *poky/chuk/ikk1* is required for differentiation of the zebrafish embryonic epidermis. *Developmental biology*, 346(2), pp. 272–283.

Fürthauer, M., Thisse, C. & Thisse, B., 1997. A role for FGF-8 in the dorsoventral patterning of the zebrafish gastrula. *Development*, 124(21), pp. 4253–4264.

Gerdes, J. M. et al., 2007. Disruption of the basal body compromises proteasomal function and perturbs intracellular Wnt response. *Nature Genetics,* 39(11), pp. 1350–1360.

Gilardelli, C.N. et al., 2004. Functional and hierarchical interactions among zebrafish *vox/vent* homeobox genes. *Developmental Dynamics*, 230(3), pp. 494–508.

Hammerschmidt, M. et al., 1996. *dino* and *mercedes*, two genes regulating dorsal development in the zebrafish embryo. *Development*, 123(1), pp. 95–102.

He, Y. et al., 2014. Maternal control of axial–paraxial mesoderm patterning via direct transcriptional repression in zebrafish. *Developmental biology*, 386(1), pp. 96–110.

Inbal, A., Topczewski, J. & Solnica-Krezel, L., 2006. Targeted gene expression in the zebrafish prechordal plate. *Genesis*, 44(12), pp. 584–588.

Jia, S. et al., 2012. Protein Phosphatase 4 Cooperates with Smads to Promote BMP Signaling in Dorsoventral Patterning of Zebrafish Embryos. *Developmental cell*, 22(5), pp. 1065–1078.

Jurynec, M.J. & Grunwald, D.J., 2010. SHIP2, a factor associated with diet-induced obesity and insulin sensitivity, attenuates FGF signaling *in vivo*. *Disease Models & Mechanisms*, 3(11-12), pp. 733–742.

Kapp, L.D. et al., 2013. The Integrator Complex Subunit 6 (Ints6) Confines the Dorsal Organizer in Vertebrate Embryogenesis G. S. Barsh, ed. *PLoS genetics*, 9(10), p. e1003822.

Kawahara, A. et al., 2000. Functional interaction of *vega2* and *goosecoid* homeobox genes in zebrafish. *Genesis*, 28(2), pp. 58–67.

Kawamura, A. et al., 2005. Zebrafish hairy/enhancer of split protein links FGF signaling to cyclic gene expression in the periodic segmentation of somites. *Genes & development*, 19(10), pp. 1156–1161.

Koshida, S. et al., 2002. Inhibition of BMP Activity by the FGF Signal Promotes Posterior Neural

Development in Zebrafish. *Developmental biology*, 244(1), pp. 9–20.

Krens, S.G. et al., 2008. ERK1 and ERK2 MAPK are key regulators of distinct gene sets in zebrafish embryogenesis. *BMC Genomics*, 9, p. 196.

Kudoh, T. et al., 2004. Combinatorial Fgf and Bmp signalling patterns the gastrula ectoderm into prospective neural and epidermal domains. *Development*, 131(15), pp. 3581–3592.

Kunwar, P.S. et al., 2003. Mixer/Bon and FoxH1/Sur have overlapping and divergent roles in Nodal signaling and mesendoderm induction. *Development*, 130(23), pp. 5589–5599.

Lele, Z., Nowak, M. & Hammerschmidt, M., 2001. Zebrafish *admp* is required to restrict the size of the organizer and to promote posterior and ventral development. *Developmental Dynamics*, 222(4), pp. 681–687.

Li, S. et al., 2010. Expression of ventral diencephalon-enriched genes in zebrafish. *Developmental Dynamics*, 239(12), pp. 3368–3379.

Lin, X. et al., 2007. Depletion of Med10 enhances Wnt and suppresses Nodal signaling during zebrafish embryogenesis. *Developmental biology*, 303(2), pp. 536–548.

Lyman Gingerich, J. et al., 2005. *hecate*, a zebrafish maternal effect gene, affects dorsal organizer induction and intracellular calcium transient frequency. *Developmental biology*, 286(2), pp. 427–439.

Maegawa, S., Varga, M. & Weinberg, E.S., 2006. FGF signaling is required for beta-catenin-mediated induction of the zebrafish organizer. *Development*, 133(16), pp. 3265–3276.

Mei, W. et al., 2009. hnRNP I is required to generate the Ca2+ signal that causes egg activation in zebrafish. *Journal of Embryology and Experimental Morphology*, 136(17), pp. 3007–3017.

Mudumana, S.P. et al., 2008. *odd skipped related 1* reveals a novel role for endoderm in regulating kidney versus vascular cell fate. *Development*, 135(20), pp. 3355–3367.

Okuda, Y. et al., 2010. B1 SOX Coordinate Cell Specification with Patterning and Morphogenesis in the Early Zebrafish Embryo. *PLoS genetics*, 6(5), p. e1000936.

Pauli, A. et al., 2014. Toddler: An Embryonic Signal That Promotes Cell Movement via Apelin Receptors. *Science*, 343(6172), pp. 1248636–1248636.

Ramel, M.-C. et al., 2005. WNT8 and BMP2B co-regulate non-axial mesoderm patterning during zebrafish gastrulation. *Developmental biology*, 287(2), pp. 237–248.

Reim, G. & Brand, M., 2006. Maternal control of vertebrate dorsoventral axis formation and epiboly by the POU domain protein Spg/Pou2/Oct4. *Development*, 133(14), pp. 2757–2770.

Reiter, J.F., Verkade, H. & Stainier, D.Y.R., 2001. Bmp2b and Oep Promote Early Myocardial Differentiation through Their Regulation of gata5. *Developmental biology*, 234(2), pp. 330–338.

Scott, I.C. et al., 2007. The *g* protein-coupled receptor *agtrl1b* regulates early development of myocardial progenitors. *Developmental cell*, 12(3), pp. 403–413.

Seo, J. et al., 2010. Negative regulation of *wnt11* expression by Jnk signaling during zebrafish gastrulation. *Journal of cellular biochemistry*, 110(4), pp. 1022–1037.

Shimizu, T. et al., 2005. Interaction of Wnt and *caudal*-related genes in zebrafish posterior body formation. *Developmental biology*, 279(1), pp. 125–141.

Sidi, S. et al., 2003. Maternal induction of ventral fate by zebrafish *radar*. *Proceedings of the National Academy of Sciences of the United States of America*, 100(6), pp. 3315–3320.

Thisse, B. & Thisse, C., 2004. *Fast Release Clones: A High Throughput Expression Analysis*, ZFIN Direct Data Submission.

Thisse, B. & Thisse, C., 2005. *High Throughput Expression Analysis of ZF-Models Consortium Clones*, ZFIN Direct Data Submission.

Thisse, B. et al., 2001. *Expression of the zebrafish genome during embryogenesis*, ZFIN Direct Data Submission.

Tian, J. et al., 2003. A temperature-sensitive mutation in the nodal-related gene cyclops reveals that the floor plate is induced during gastrulation in zebrafish. *Development*, 130(14), pp. 3331–3342.

Tian, J. et al., 2008. The pro-domain of the zebrafish Nodal-related protein Cyclops regulates its signaling activities. *Development*, 135(15), pp. 2649–2658.

Toyama, R. & Dawid, I.B., 1997. *lim6*, a novel LIM homeobox gene in the zebrafish: Comparison of its expression pattern with *lim1*. *Developmental Dynamics*, 209(4), pp. 406–417.

Tseng, W.-F. et al., 2011. An evolutionarily conserved kernel of *gata5*, *gata6*, *otx2* and *prdm1a* operates in the formation of endoderm in zebrafish. *Developmental biology*, 357(2), pp. 541–557.

Walshe, J. et al., 2002. Establishment of Hindbrain Segmental Identity Requires Signaling by FGF3 and FGF8. *Current Biology*, 12(13), pp. 1117–1123.

Wang, W.-D. et al., 2011. Tfap2a and Foxd3 regulate early steps in the development of the neural crest progenitor population. *Developmental biology*, 360(1), pp. 173–185.

Warga, R.M. & Kane, D.A., 2003. One-eyed pinhead regulates cell motility independent of Squint/Cyclops signaling. *Developmental biology*, 261(2), pp. 391–411.

Wilm, T.P. & Solnica-Krezel, L., 2005. Essential roles of a zebrafish *prdm1/blimp1* homolog in embryo patterning and organogenesis. *Development*, 132(2), pp. 393–404.

Wu, S.-Y. et al., 2012. Chemokine GPCR Signaling Inhibits β-Catenin during Zebrafish Axis Formation M. C. Mullins, ed. *PLoS biology*, 10(10), p. e1001403.

Xie, X.-W. et al., 2011. Zebrafish *foxo3b* Negatively Regulates Canonical Wnt Signaling to Affect Early Embryogenesis Z. Wen, ed. *PloS one*, 6(9), p. e24469.

Yao, S. et al., 2010. Kzp controls canonical Wnt8 signaling to modulate dorsoventral patterning during zebrafish gastrulation. *The Journal of biological chemistry*, 285(53), pp. 42086–42096.

Yimlamai, D. et al., 2005. The zebrafish *down syndrome cell adhesion molecule* is involved in cell movement during embryogenesis. *Developmental biology*, 279(1), pp. 44–57.

Zhao, J. et al., 2003. An SP1-like transcription factor Spr2 acts downstream of Fgf signaling to mediate mesoderm induction. *The EMBO journal*, 22(22), pp. 6078–6088.

Zon, L., Hammerschmidt, M. & Schulte-Merker, S., 1997. The molecular nature of zebrafish *swirl*: BMP2 function is essential during early dorsoventral patterning. *Development*, 124(22), pp. 4457–4466.

SUPPLEMENTARY TABLE 2
Generation of *in situ* probes

Gene	Forward Primer	Reverse Primer	Linearize	Polymerase
aplnrb			BamHI	T3
arl4ab	attggactgtgccggtaaga	aacatttgtgcgttccccaa	NotI	T3
casp8	acaattggccgcattgactt	gtcggtaggagaggtagtgc	NotI	G3
cpn1	catttacagcatcggtcgca	tgtaggtacccggcaaaagt	NotI	T3
dusp4	actgcagcgttttgaagagg	gaaacgcgcccttactagtg	EcoRV	T7
ets2			Acc65I	SP6
gadd45aa	gacagccagagaaagaacacc	ggtgctcacttccattcaca	NotI	T3
id2a	gcgaacagggaatctcgaac	aacactttgcagataccggc	NotI	T3
igf2a	gaggaatgctgctttcggag	tgggcctacttgattgcaga	EcoRV	T7
insm1b	gccagtcagcaaggatcatg	cttcagcaggcgttacgtac	EcoRV	T7
irx7	cttcatcaacggggtttgca	acaggagagtacgagtccct	NotI	T3
isg15	tcatcacagttgttggcaca	acatcacggcattgaaaacac	EcoRV	T7
nrarpa	tgcagaacatgaccaactgc	atcggttgctttctccagga	NotI	T3
pkdcca	aatggacctggagcaacgta	actctgttttgtgcatgcgt	NotI	T3
prdm12b	gttcggctcatcatgggttc	gtggtgcttgatcccaatgg	Acc65I	T7
prickle1b	tgacatgtattgggcccagt	aactttgggatcggggctta	NotI	T3
ripply1	gttttctcaccctcacgctg	gcgatgcggtgttcaatgtta	NotI	T3
slc25a33	gttttccaggttcagctggg	tggtcagactcccaaatgct	NotI	T3
tbr1b	ccatgtttccctacccgagt	gcaaagtccagtcggttgtt	NotI	T3
tcf3b	gtggatggaggtggtcaaga	ccattgacgtctgctccatg	NotI	T3
tp53inp1	tggttcatcactcctccacc	agaatcacaggcaggttcca	EcoRV	T7

SUPPLEMENTARY NOTE:

Guided Seurat analysis of Zebrafish dataset

Seurat

Rahul Satija and Jeffrey Farrell
rahuls@broadinstitute.org and jfarrell@g.harvard.edu

Installation:
Extract all files from this package to a single directory.

Requirements:
(1) A working installation of **R** (http://cran.r-project.org/) and **RStudio** (http://www.rstudio.com/) with **R Markdown** installed (RStudio will offer to install this for you).

(2) The **following packages** are required by Seurat (and can be installed from RStudio using the *install.packages* command): **vioplot, reshape2, XLConnect, lars, mixtools, NMF, gplots, reshape, Hmisc, ggplot, ROCR, gdata, and rgl**.

(3) For Seurat_AnalyzePopulations_D, a **working installation of X11** is required. This is absolutely critical, as an attempt to load the rgl package without X11 installed does not fail gracefully and will crash R. On Macs, we recommend the XQuartz package (http://xquartz.macosforge.org/).

Use:
The R Markdown files reproduce most of the computation analysis described in the manuscript. They are a series of 4 modular files that save their computed output as R objects that are loaded in by the next module. This allows users to experiment with each module without having to run the (computationally intensive) analysis from scratch each time. The files have been knit into PDFs that show the output of running each module. The parts perform the following analyses:

(1) Load in and normalize expression data, identify EVL cells
(2) PCA analysis to identify variable and structured genes, building models of gene expression to alleviate technical noise
(3) Probabilistic inference of spatial origin : fitting Gaussian mixture models, additional quantitative refinement, and cell projection.
(4) Guided dataset analysis. Inference of *in silico in situ* patterns generated from the projected cells for both known and novel spatial markers, combining unsupervised analyses with Seurat to identify and localize rare populations.

Seurat - Load in Data and Identify EVL

Rahul Satija and Jeff Farrell

September 8, 2014

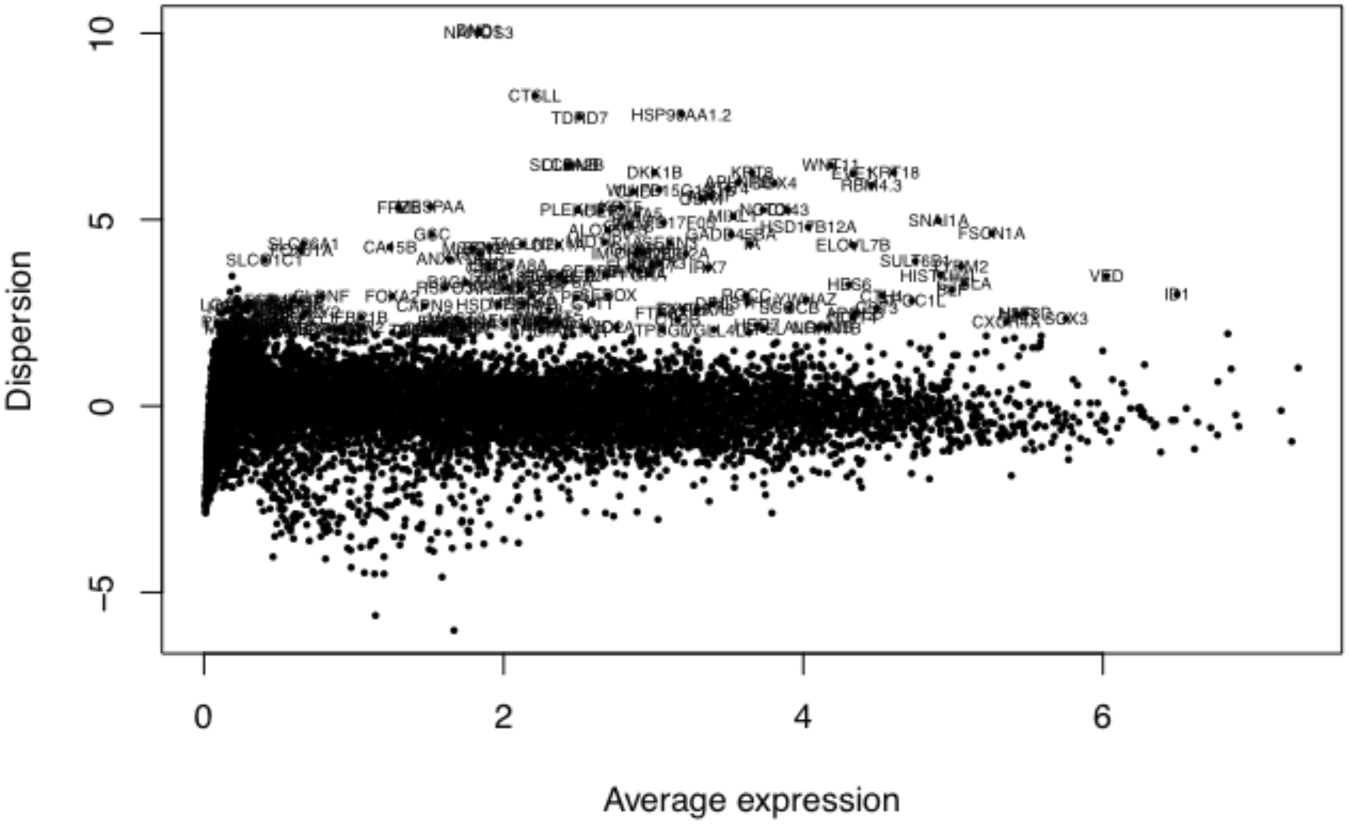
Set basic options and load requirements Note the following package requirements: vioplot,reshape2,XLConnect,lars,mixtools,NMF,gplots,reshape,Hmisc,ggplot2,ROCR

Load-in and normalize data

```
filesToLoad=c("~/seurat/data/zfish_umi_counts.Robj")
for(loadFile in filesToLoad) load(loadFile)

zdata.norm.slim=log(sweep(zdata,2,colSums(zdata),"/")*2e5+1)
zdata.norm.slim=minusr(zdata.norm.slim,"SI:|ORF|^CU|^ZGC|^BX");

zf.all=seurat(raw.data=zdata.norm.slim,stat.fxn=getStat1,is.expr=0.01)
zf.all=setup(zf.all,project="Seurat",min.cells = 3,min.genes = 2000,calc.noise=FALSE,is.exp
r=0.01,do.scale = TRUE) #take all genes in > 3 cells, all cells with > 2k genes
```

Identify variable genes across the single cells

```
zf.all=mean.var.plot(zf.all,y.cutoff = 2,do.plot=TRUE,x.low.cutoff=0.25,x.high.cutoff=7,fxn
.x = expMean,fxn.y=logVarDivMean,set.var.genes = TRUE)
```



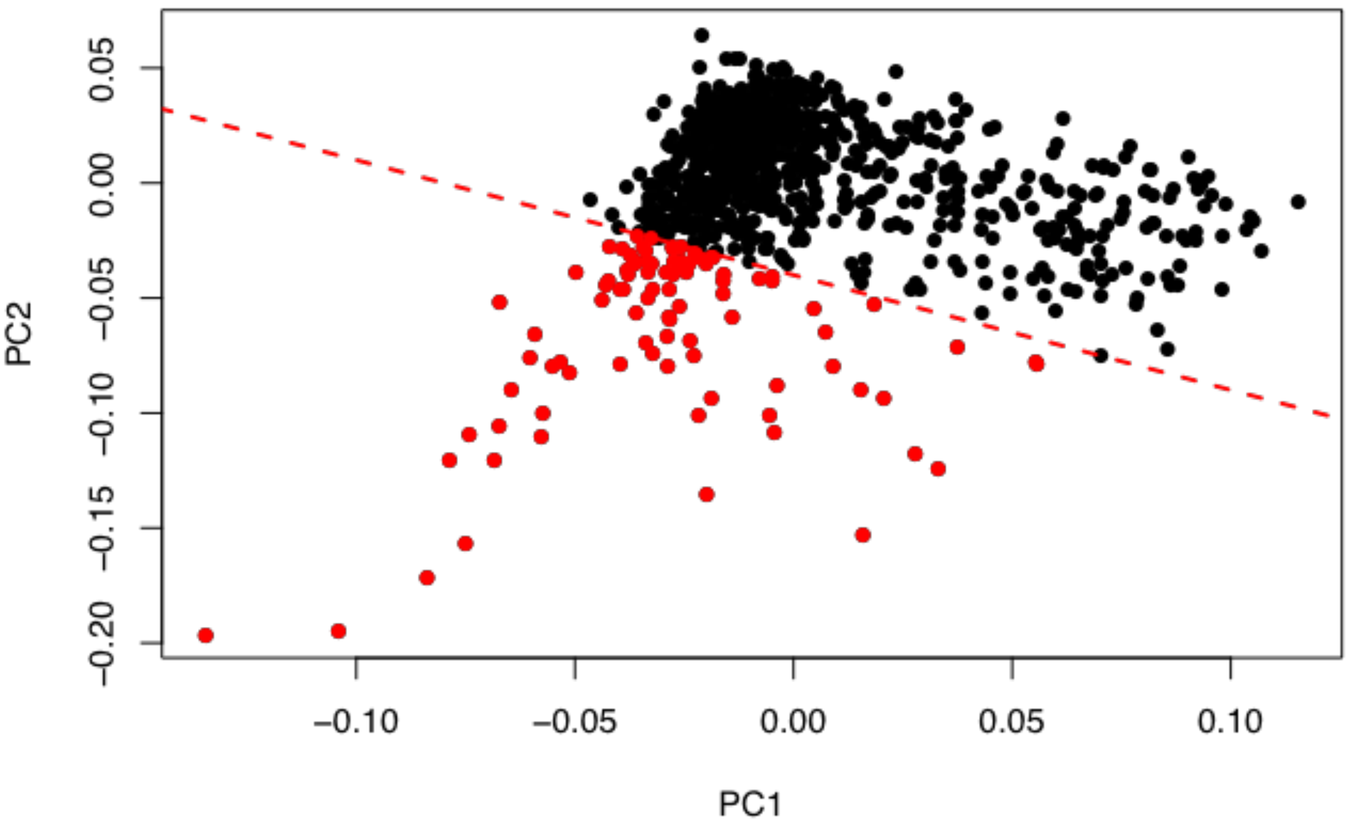
```
markers.remove=batch.gene(zf.all,stats.use = c("zf1","zf2","zf3"),genes.use=zf.all@var.gene
s,auc.cutoff = 0.7)
zf.all@var.genes=zf.all@var.genes[!(zf.all@var.genes%in%markers.remove)]
```

Run a PCA using only these variable genes, identify EVL cells

```
zf.all=pca(zf.all,do.print = FALSE)
zf.all=project.pca(zf.all, pcs.print = 2,genes.print = 8)
```

```
## [1] "PC1"
## [1] "SOX3"     "SOX19A"   "CXCR4B"   "ID1"      "ALCAMB"   "ZIC2B"
## [7] "FAM212AA" "GLULB"
## [1] ""
## [1] "OSR1"     "MIXL1"    "DKK1B"    "KIRREL3L" "ZIC2A"    "APLNRB"
## [7] "GATA5"    "MESPAB"   "FLRT3"
## [1] ""
## [1] ""
## [1] "PC2"
## [1] "KRT18"      "KRT4"       "WU:FB15G10" "KRT8"       "CEBPB"
## [6] "WU:FB17F05" "KRT5"       "CLDNE"
## [1] ""
## [1] "CXCR4B" "ZIC2B"  "ID1"    "SOX3"   "ALDOB"  "CXCR4A" "SOX19A" "ANP32A"
## [9] "DKC1"
## [1] ""
## [1] ""
```

```
plot(zf.all@pca.rot[,1],zf.all@pca.rot[,2],pch=16,xlab="PC1",ylab="PC2")
x=seq(-0.2,0.2,.01)
lines(x,-x*0.5-0.04,lwd=2,lty=2,col="red")
evl.quant=zf.all@pca.rot[,1]+2*zf.all@pca.rot[,2]+0.08; names(evl.quant)=colnames(zf.all@data)
not.evl=names(evl.quant[evl.quant>0])
is.evl=names(evl.quant[evl.quant<0])
points(zf.all@pca.rot[is.evl,1],zf.all@pca.rot[is.evl,2],pch=16,col="red")
```



Save the data so we can move to the next RMD without having to Reload the data

```
save(zf.all,not.evl,is.evl,file="~/seurat/obj/output_A.Robj")
```

Seurat_ImputeData_B

Rahul Satija and Jeff Farrell

September 8, 2014

Set basic options and load requirements Note the following package requirements: vioplot,reshape2,XLConnect,lars,mixtools,NMF,gplots,reshape,Hmisc,ggplot2,ROCR

```
#loads zf object, not.evl, is.evl
filesToLoad=c("~/seurat/obj/output_A.Robj")
for(loadFile in filesToLoad) load(loadFile)

#remove the EVL cells by taking a subset of the data
zf <- subsetData(zf.all, "", cells.use=not.evl)
```

Identify genes to use for building gene expression models

```
#recalculate a set of variable genes
zf <- mean.var.plot(zf, y.cutoff = 2, do.plot=FALSE, x.low.cutoff=1, x.high.cutoff=7, fxn.x = expMean, fxn.y=logVarDivMean, set.var.genes = TRUE)
markers.remove=batch.gene(zf,stats.use = c("zf1","zf2","zf3"),genes.use=zf@var.genes,auc.cutoff = 0.6)
zf@var.genes=zf@var.genes[!(zf@var.genes%in%markers.remove)]

#do PCA to identify 'structured' genes
#note that the PCA prints the 8 genes with the highest and lowest loadings for each PC
zf <- pca(zf, do.print = FALSE)
zf <- jackStraw(zf, num.replicate=1000, num.pc=4, prop.freq=0.025)
zf <- project.pca(zf, pcs.print = 4, genes.print = 8)
```

```
## [1] "PC1"
## [1] "SOX3"     "SOX19A"   "CXCR4B"   "ID1"      "ZIC2B"    "ALDOB"
## [7] "CITED4B"  "FAM212AA"
## [1] ""
## [1] "OSR1"     "MIXL1"    "GATA5"    "DKK1B"    "KIRREL3L" "APLNRB"
## [7] "MESPAB"   "ZIC2A"    "FLRT3"
## [1] ""
## [1] ""
## [1] "PC2"
## [1] "CHD"   "FRZB"  "GSC"   "ADMP"  "FOXA2" "OTX1A" "SIX3B" "FOXA3"
## [1] ""
## [1] "BAMBIA" "VED"    "VOX"    "EVE1"   "APOC1L" "VENT"   "WNT8A"  "WNT8-2"
## [9] "CDX4"
## [1] ""
## [1] ""
## [1] "PC3"
## [1] "IRX7"   "OTX1A"  "SHISA2" "LFNG"   "FGFR4"  "RND1L"  "PITX2"  "LFT2"
## [1] ""
## [1] "NOTO"    "FOXD5"   "ARL4AB"  "WNT11"   "TA"      "AMOTL2A" "ARG2"
## [8] "HES6"    "DUSP6"
## [1] ""
## [1] ""
## [1] "PC4"
## [1] "CXCR4A"   "DND1"     "SCRN2"    "SOX32"    "OTX1A"    "ATP6V1G1"
## [7] "OTX1B"    "GRA"
## [1] ""
## [1] "ISG15"  "SESN3"  "PHLDA3" "CTSH"   "MAT2AL" "KRT18"  "FOXO3B" "PHLDA2"
## [9] "IGF2A"
## [1] ""
## [1] ""
```

Build models of gene expression

Matrices of gene expression were generated from published in situ stainings, and saved in an Excel file (which eases data entry). So, we import this data and add it to the Seurat object.

```
# Load in the Excel file.
wb <- loadWorkbook("~/seurat/situ/Spatial_ReferenceMap.xlsx", create=FALSE); insitu.genes <
- getSheets(wb)
insitu.matrix <- data.frame(sapply(1:length(insitu.genes),function(x)as.numeric(as.matrix(w
b[x][2:9,2:9]))))
insitu.genes <- toupper(insitu.genes); colnames(insitu.matrix) <- (insitu.genes)

# Then, we store this information in the Seurat object.
zf@insitu.matrix=insitu.matrix[,insitu.genes]
```

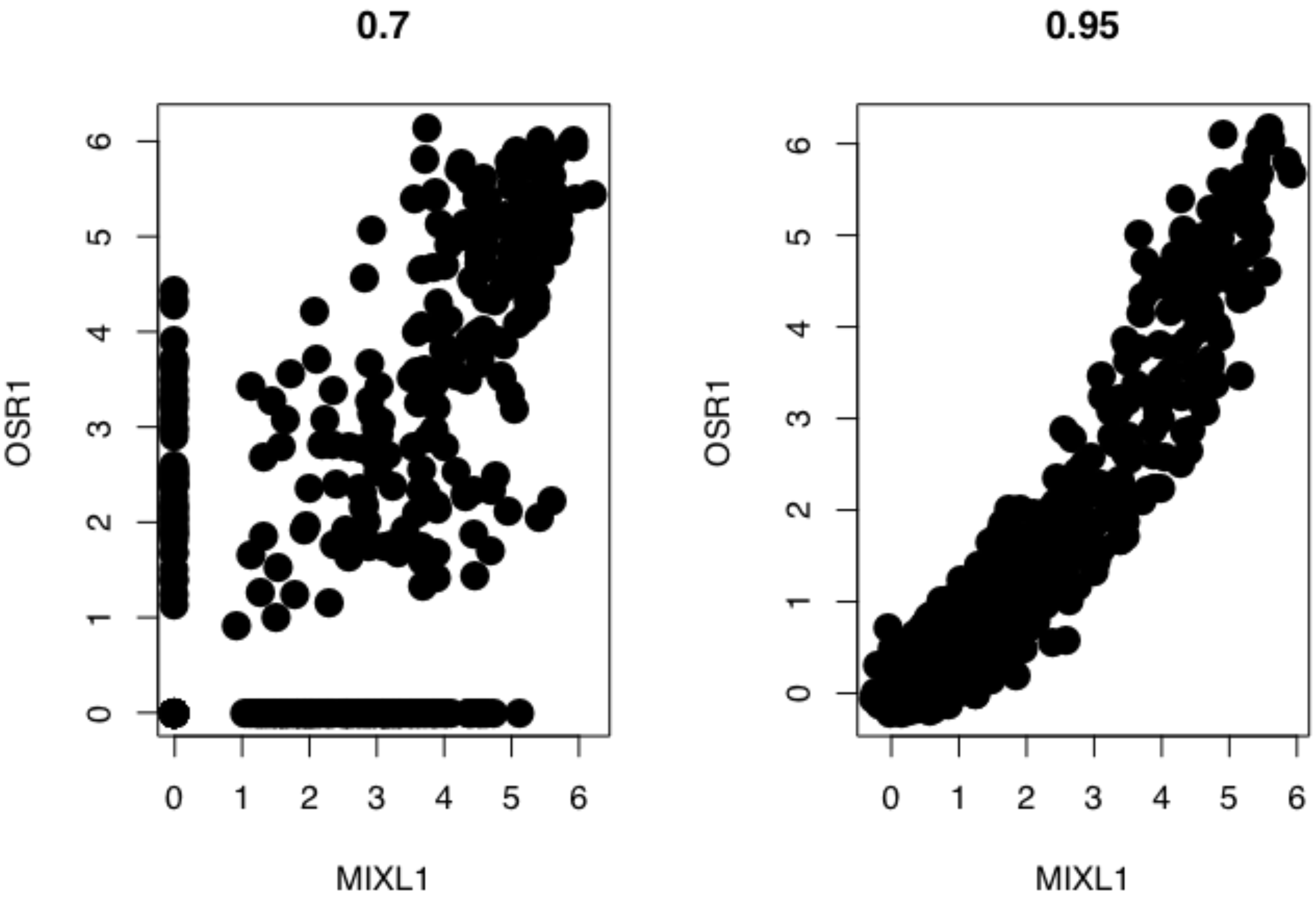
Now build models for these insitu genes, and predict robust values

```
genes.sig <- pca.sig.genes(zf,pcs.use = c(1,2,3), pval.cut = 1e-2, use.full = TRUE)
lasso.genes.use=unique(c(genes.sig,zf@var.genes))
zf <- addImputedScore(zf, genes.use=lasso.genes.use,genes.fit=insitu.genes, do.print=FALSE,
 s.use=40, gram=FALSE)
```

```
##  [1] "ADMP"   "APLNRB" "AXIN2"  "BAMBIA" "BMP2B"  "BMP4"   "BMP7A"
##  [8] "CDX4"   "CHD"    "CST3"   "EVE1"   "FGF8A"  "FOXB1A" "FOXD3"
## [15] "FOXD5"  "FOXI1"  "GATA2A" "GATA5"  "GSC"    "HER1"   "HES6"
## [22] "ID3"    "ISM1"   "LHX1A"  "LFT1"   "LFT2"   "MIXL1"  "NDR1"
## [29] "NDR2"   "NOG1"   "NOTO"   "OSR1"   "OTX1A"  "OTX1B"  "SEBOX"
## [36] "SNAI1A" "SOX3"   "SP5L"   "SZL"    "TA"     "TBX16"  "TPH1B"
## [43] "VED"    "VENT"   "VOX"    "WNT8A"  "WNT11"
```

Demonstrate the benefit of imputation

```
#before imputation - MIXL1 and OSR1 should be tightly co-expressed (on the left)
par(mfrow=c(1,2))
genePlot(zf,"MIXL1","OSR1",col="black")
#after imputation (on the right)
genePlot(zf,"MIXL1","OSR1",use.imputed = TRUE,col="black")
```



Save the data so we can move to the next RMD without having to Reload the data

```
save(zf,lasso.genes.use,file="~/seurat/obj/output_B.Robj")
```

Seurat_InferOrigins_C

Rahul Satija and Jeff Farrell

September 8, 2014

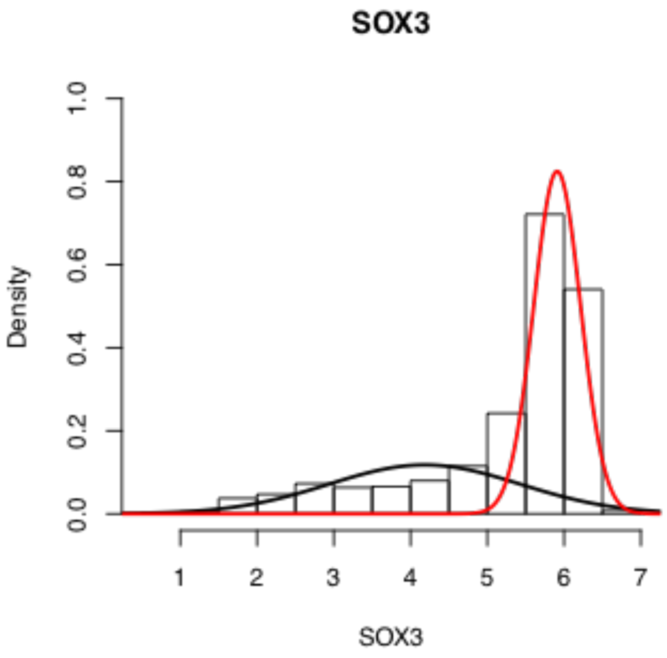
Set basic options and load requirements Note the following package requirements: vioplot,reshape2,XLConnect,lars,mixtools,NMF,gplots,reshape,Hmisc,ggplot2,ROCR

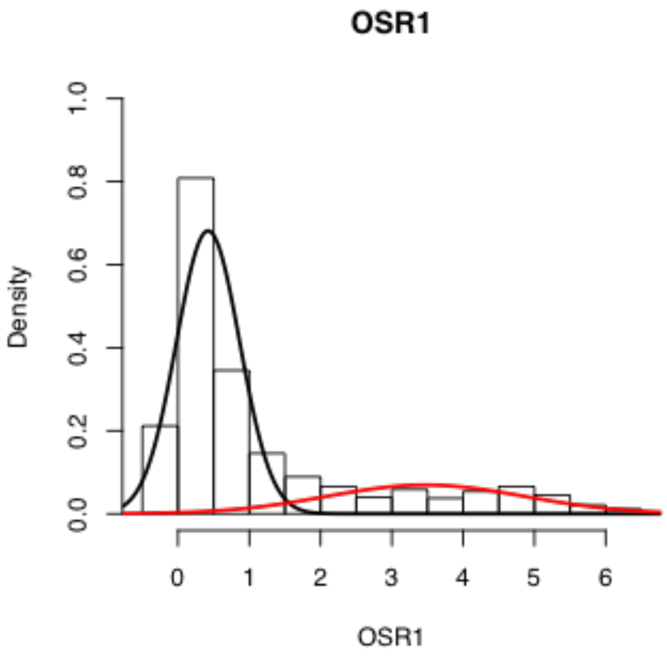
```
#loads zf object, not.evl, is.evl
filesToLoad=c("~/seurat/obj/output_B.Robj")
for(loadFile in filesToLoad) load(loadFile)
```

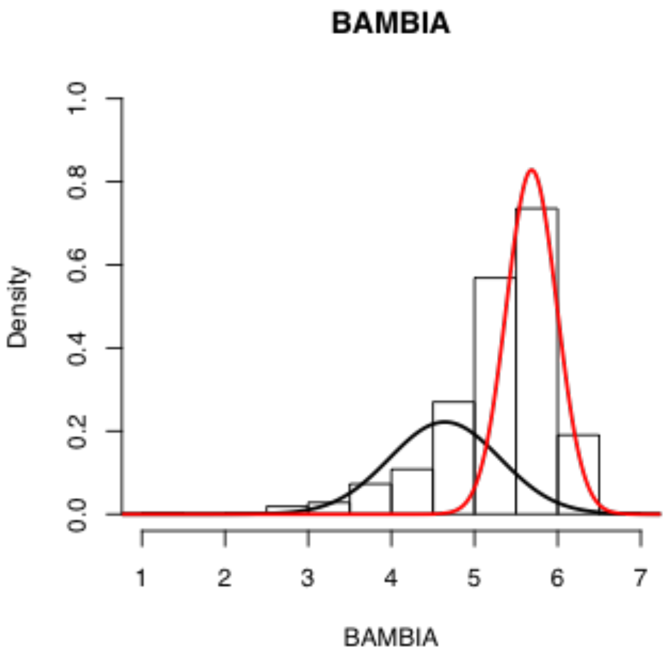
Mixture models

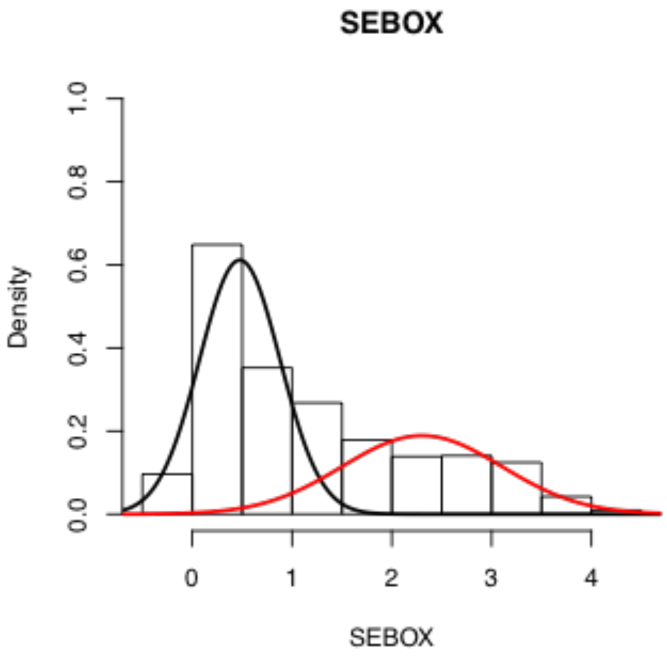
The in situ patterns that we use to provide geographical information are scored in a binary on/off format. In order to translate the continuous RNAseq data into this form, we model it as mixtures of 2 normal distributions that represent the on state and off state. We then use this to estimate whether each cell should be considered on or off for each gene.

```
insitu.genes=colnames(zf@insitu.matrix)
for(i in rev(insitu.genes)) zf=fit.gene.k(zf,i,do.plot=FALSE,do.k = 2,start.pct=mean(zf@insitu.matrix[,i]),num.iter = 1)

#show an example mixture model
par(mfrow=c(2,2))
zf_temp=fit.gene.k(zf,"SOX3",do.plot=TRUE,do.k = 2,start.pct=mean(zf@insitu.matrix[,"SOX3"]),num.iter = 1)
zf_temp=fit.gene.k(zf,"OSR1",do.plot=TRUE,do.k = 2,start.pct=mean(zf@insitu.matrix[,"OSR1"]),num.iter = 1)
zf_temp=fit.gene.k(zf,"BAMBIA",do.plot=TRUE,do.k = 2,start.pct=mean(zf@insitu.matrix[,"BAMBIA"]),num.iter = 1)
zf_temp=fit.gene.k(zf,"SEBOX",do.plot=TRUE,do.k = 2,start.pct=mean(zf@insitu.matrix[,"SEBOX"]),num.iter = 1)
```









```
# Project each cell into its proper location
zf <- initial.mapping(zf)
```

Now, do a quantitative refinement

```
#first identify the genes to use for the refinement
num.pc=3; num.genes=3;
genes.high=as.vector(apply(zf@pca.x.full[,1:num.pc],2,function(x)rownames(zf@pca.x.full)[order(x)[1:num.genes]]))
genes.low=as.vector(apply(zf@pca.x.full[,1:num.pc],2,function(x)rownames(zf@pca.x.full)[order(x,decreasing=TRUE)[1:num.genes]]))
genes.use=unique(c(genes.high,genes.low))

#impute values for these genes if needed
new.imputed=genes.use[!genes.use%in%rownames(zf@imputed)]
zf <- addImputedScore(zf, genes.use=lasso.genes.use,genes.fit=new.imputed, do.print=FALSE, s.use=40, gram=FALSE)
```

```
## [1] "SOX19A" "CXCR4B" "FRZB"   "IRX7"   "SHISA2" "ARL4AB"
```

```
#refine the mapping with quantitative models that also consider gene covariance
zf <- refined.mapping(zf,genes.use)
```

Save the data so we can move to the next RMD without having to Reload the data

```
save(zf,lasso.genes.use,file="~/seurat/obj/output_C.Robj")
```

Seurat_AnalyzePopulations_D

Rahul Satija and Jeff Farrell

September 8, 2014

Set basic options and load requirements Note the following package requirements: vioplot,reshape2,XLConnect,lars,mixtools,NMF,gplots,reshape,Hmisc,ggplot2,ROCR

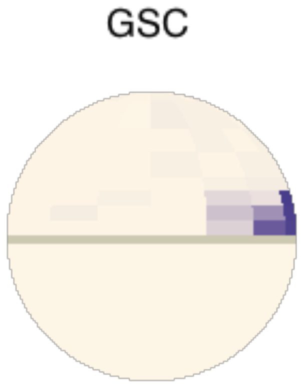
FOR ALL 3D PLOTS WITH RGL, PLEASE MAKE SURE AN X11 client (i.e. Xquartz for Mac OS X) before attempting to run this file!!

```
#loads zf object, not.evl, is.evl
filesToLoad=c("~/seurat/obj/output_C.Robj")
for(loadFile in filesToLoad) load(loadFile)
```

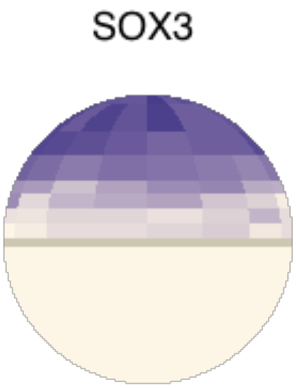
Draw inferred in situ patterns for a few known genes

Note that in the manuscript, when drawing an in situ for a ‘landmark’ gene, we first removed the gene from the spatial reference map (i.e., removed the column from zf@insitu.matrix (mailto:zf@insitu.matrix)), re-mapped the cells, and then inferred an in situ pattern

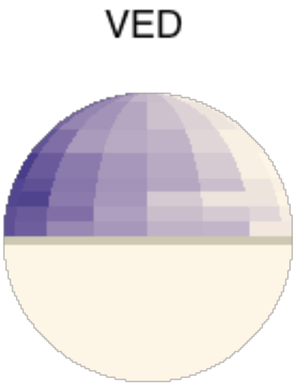
```
zf.insitu.lateral(zf, "GSC",label=FALSE)
```



```
zf.insitu.lateral(zf, "SOX3",label=FALSE)
```

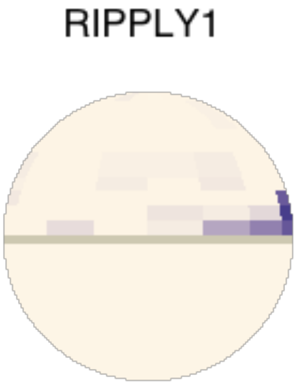


```
zf.insitu.lateral(zf, "VED",label=FALSE)
```

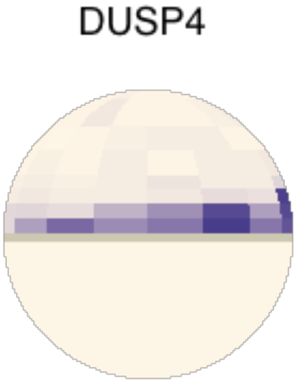


Draw inferred in situ patterns for a few new genes

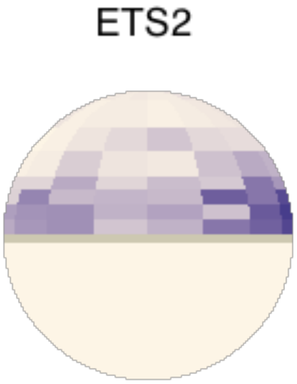
```
zf.insitu.lateral(zf, "RIPPLY1",label=FALSE)
```



```
zf.insitu.lateral(zf, "DUSP4",label=FALSE)
```



```
zf.insitu.lateral(zf, "ETS2",label=FALSE)
```

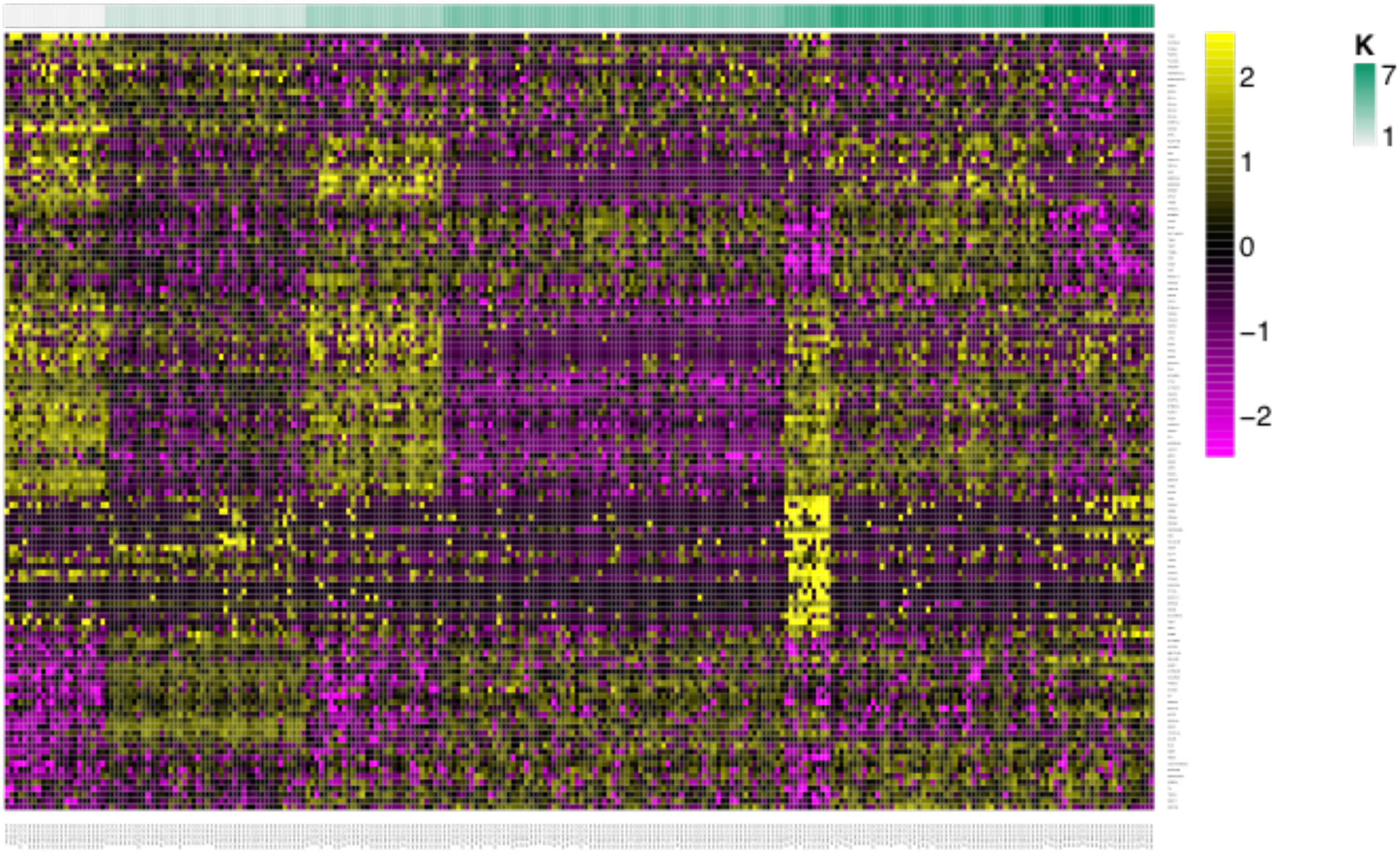


Pull out the margin cells, and run a PCA only on these cells

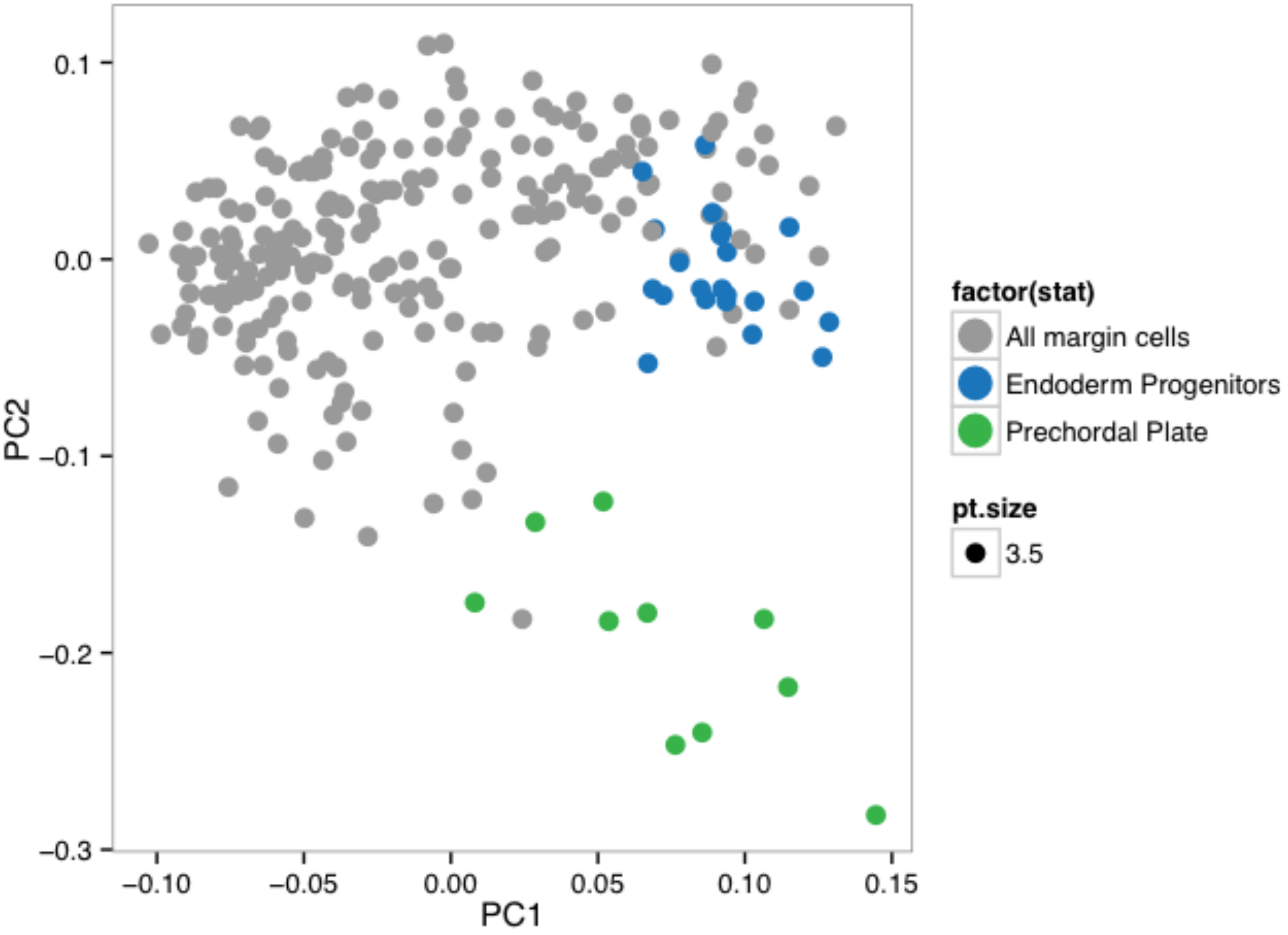
```
cell.centroids=data.frame(t(sapply(colnames(zf@data),function(x) exact.cell.centroid(zf@final.prob[,x])))); colnames(cell.centroids)=c("bin.tier","bin.dv")
margin.cells=rownames(subset(cell.centroids,bin.tier>=5))
zf.margin=(subsetData(zf,cells.use = margin.cells))
zf.margin=pca(zf.margin,do.print = FALSE)
zf.margin <- jackStraw(zf.margin, num.replicate=1000, prop.freq=0.025)
zf.margin=project.pca(zf.margin,pcs.print = 3,genes.print = 8)
```

```
## [1] "PC1"
## [1] "ID1"    "SOX3"   "SOX19A" "CXCR4B" "ZIC2B"  "FOXD5"  "ALDOB"  "ASB11" 
## [1] ""
## [1] "OSR1"     "GATA5"    "APLNRB"   "FSCN1A"   "SNAI1A"   "DKK1B"   
## [7] "PITX2"    "KIRREL3L" "ZIC2A"   
## [1] ""
## [1] ""
## [1] "PC2"
## [1] "CHD"   "GSC"   "FRZB"  "OTX1A" "NOG1"  "SIX3B" "FOXA2" "TBX1" 
## [1] ""
## [1] "VED"    "HES6"   "VOX"    "EVE1"   "BAMBIA" "WNT8A"  "WNT8-2" "TPBGL" 
## [9] "VENT"  
## [1] ""
## [1] ""
## [1] "PC3"
## [1] "SOX32"     "MARCKSL1B" "CXCR4A"    "FGFR4"     "RPL26"     "BAMBIA"   
## [7] "CPN1"      "FLOT2A"   
## [1] ""
## [1] "ARL4AB"       "WNT11"        "SEBOX"        "GADD45BA"    
## [5] "TA"           "MIXL1"        "MORC3B"       "LOC100536023"
## [9] "PRICKLE1B"   
## [1] ""
## [1] ""
```

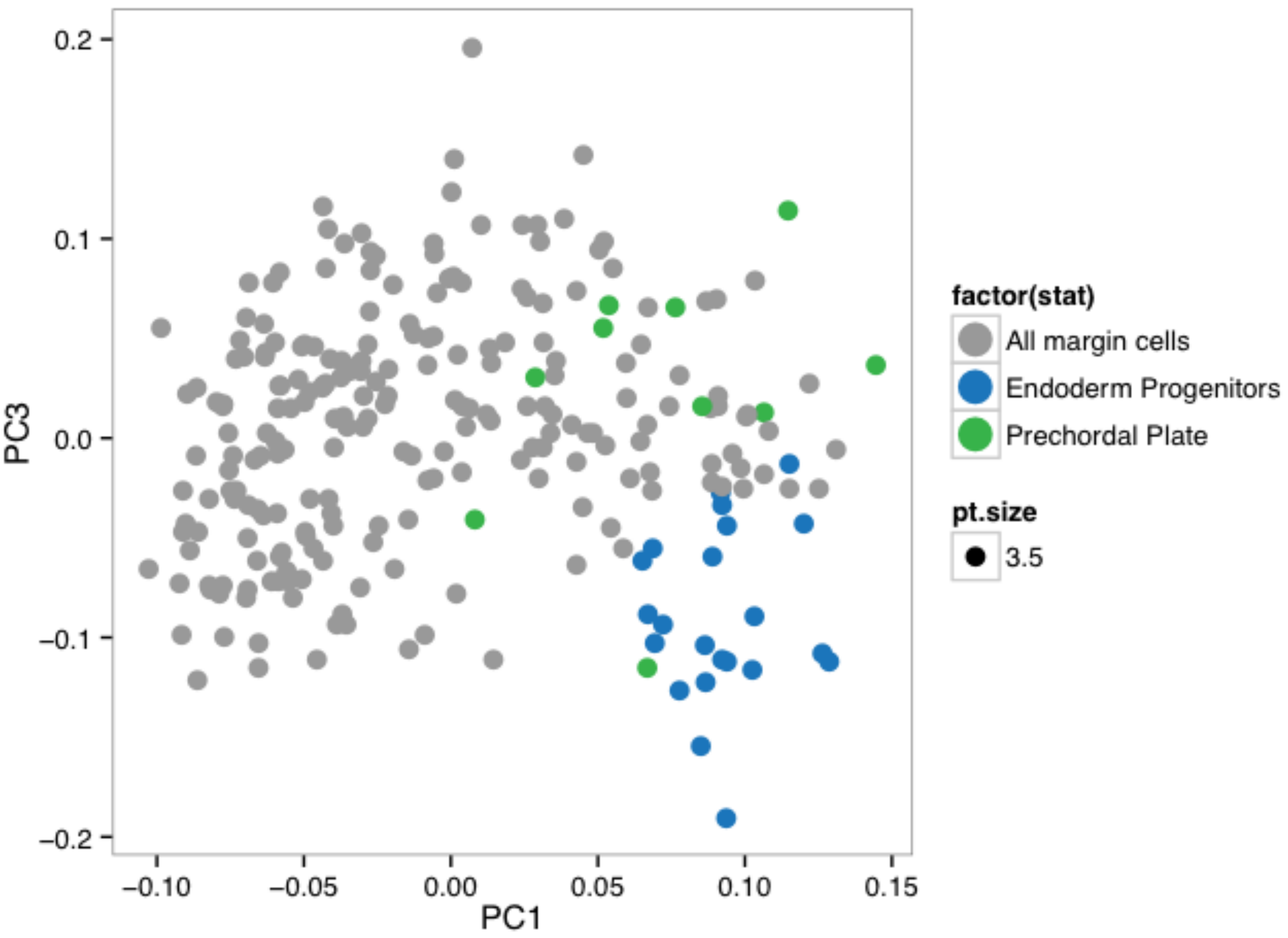
```
zf.margin=doKMeans(zf.margin,pcs.use = c(1:3), pval.cut = 1e-3,k.num = 8,k.seed = 1,disp.cut = 2.5,do.k.col = TRUE,use.full = TRUE, k.col=7,rev.pc.order = FALSE,pc.row.order = 3,pc.col.order = 3,do.plot=TRUE,print.genes=FALSE)
```



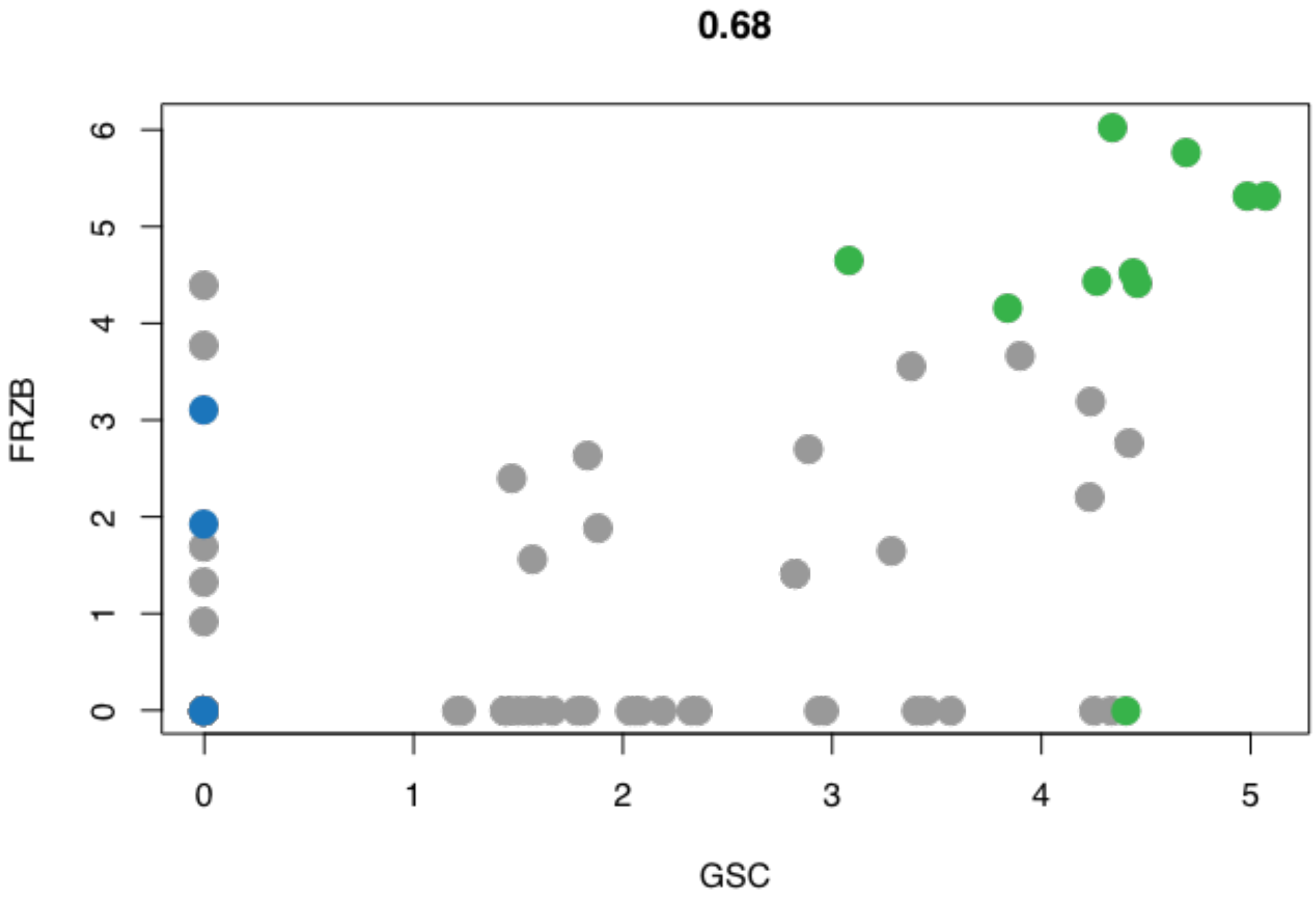
```
endo.cells=cells.in.cluster(zf.margin,1)
plate.cells=cells.in.cluster(zf.margin,5)

zf.margin@data.stat[colnames(zf.margin@data)]=rep("All",ncol(zf.margin@data));
zf.margin@data.stat[margin.cells]="All margin cells"
zf.margin@data.stat[endo.cells]="Endoderm Progenitors"; zf.margin@data.stat[plate.cells]="Prechordal Plate";
pop.cols=c("#999999","#1B75BB","#37B34A")

#plot PCA, coloring cells by their status/ID (stored in zf@data.stat)
pca.plot(zf.margin,1,2,pt.size = 3.5,cols.use = pop.cols);
```



```
pc.13=pca.plot(zf.margin,1,3,pt.size = 3.5,cols.use = pop.cols)
```

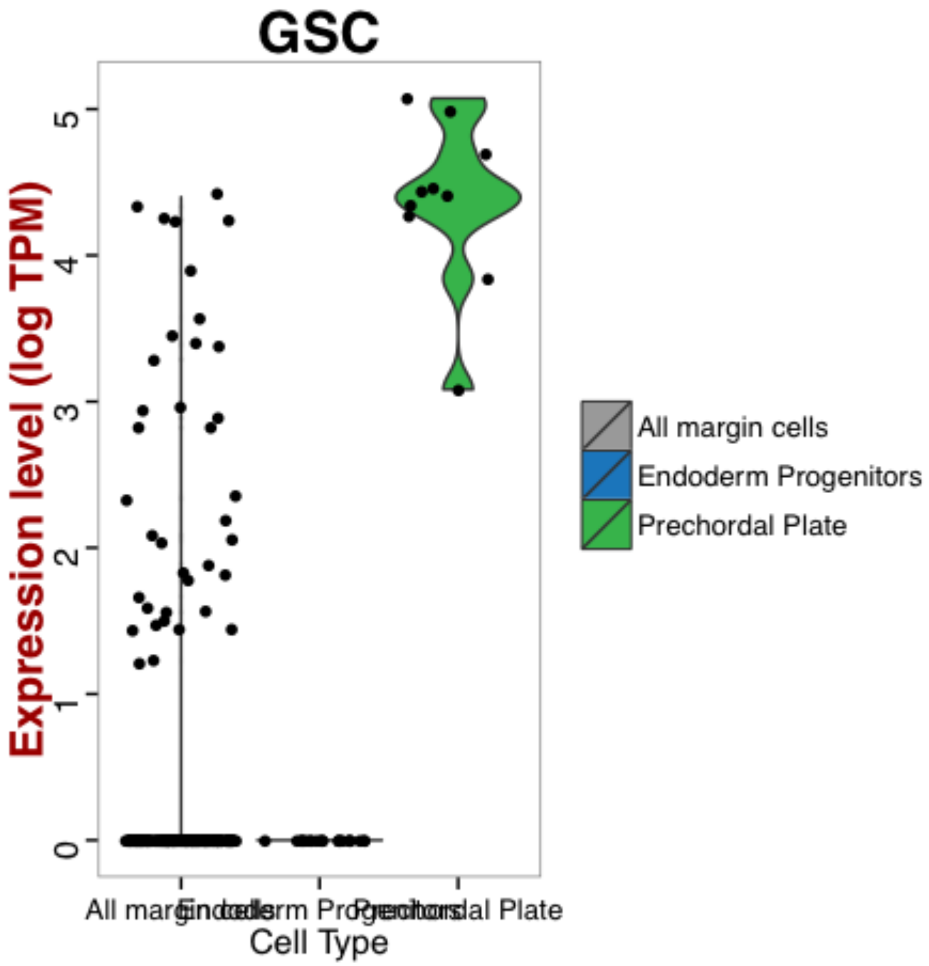


```
#gene-gene plot, again, coloring cells by their stat/ID
genePlot(zf.margin,"GSC","FRZB",col.use=pop.cols)
```

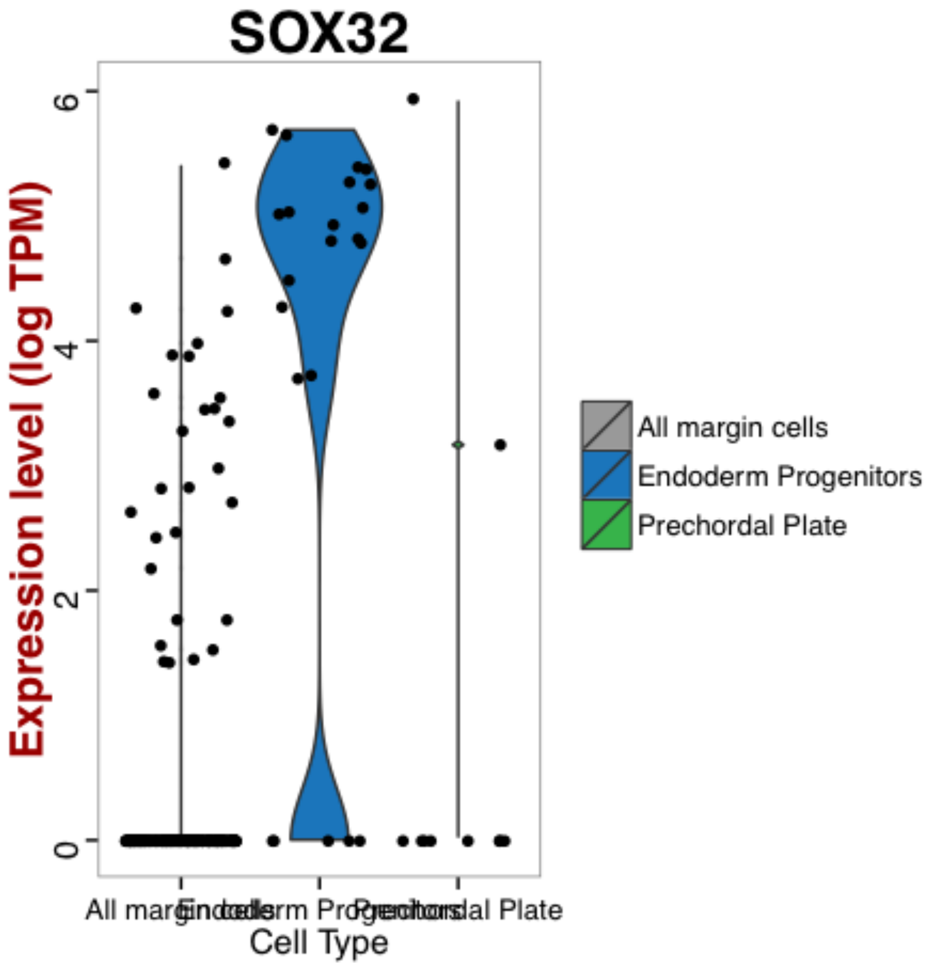


Draw violin plots of known and new markers

```
vlnPlot(zf.margin,c("GSC"),cols.use = pop.cols)
```

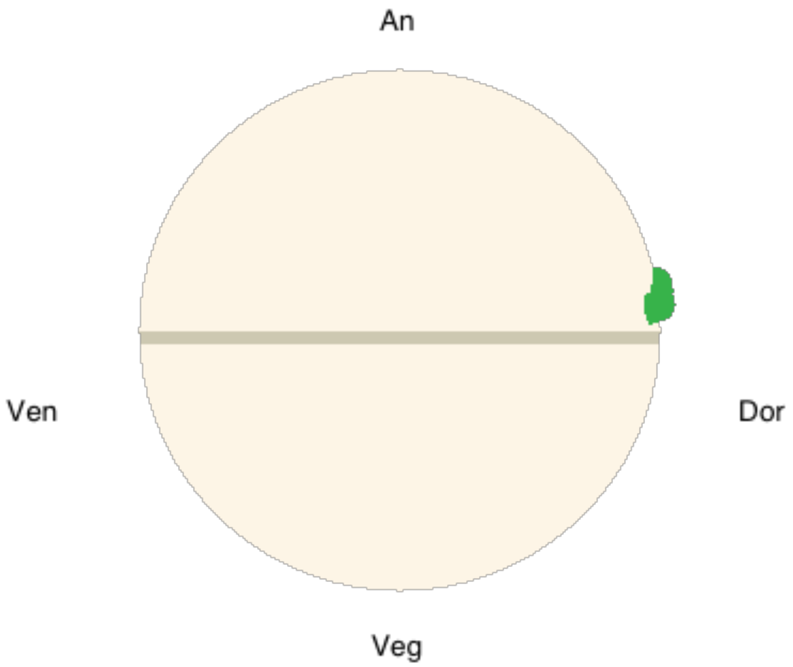


```
vlnPlot(zf.margin,c("SOX32"),cols.use = pop.cols)
```



Localize cellular populations

```
#prechordal plate
zf.cells.render(zf,plate.cells,do.rotate=FALSE,radius.use=0.0625,col.use="#37B34A",do.new=T
RUE)
```



```
#endoderm progenitors
zf.cells.render(zf,endo.cells,do.rotate=FALSE,radius.use=0.0625,col.use="#1B75BB",do.new=TR
UE,label=FALSE)
```