Supplementary Methods

Saturation analysis

To determine saturation of the libraries we numerically simulated experiments with fewer reads by randomly sampling a given fraction of total aligned reads and calculating the corresponding number of enriched regions for a given FDR- threshold. The value of FDR threshold was kept constant for all subsets. The distribution of the number of enriched regions for a constant empirical FDR-threshold of ~0.01 was then plotted as a function of total reads in the simulated experiment (Figure S16). The number of enriched regions at this FDR were linearly interpolated from the numbers estimated for the two FDR values flanking FDR 0.01, which corresponded to integer-valued height thresholds.

Seeded motif discovery

For motif discovery we used GADEM [1], which can efficiently address large sets of sequence regions. GADEM identifies conserved motifs that have *E*-values below a user-specified threshold. A motif's *E*-value is the product of its *p*-value and the number of all possible motif-length segments in the search space. When a motif has low prevalence and/or is short, particularly when the search space is large, the *E*-value can be greater than 1 and the motif can be considered not to be significant; such motifs can be difficult to identify using E-value based motif discovery tools.

The work described here involved HNF4A, FOXA2 and PDX1, whose DNA sequence motif lengths are 13 bp, 10 bp and approximately 6-7 bp (TRANSFAC M01031, [2]; [3]; and TRANSFAC M00436, respectively). Given this wide length range, we used a version of

GADEM that we modified so that motif optimization started from an initial ('seed') position weight matrix (PWM) that was not generated from a spaced dyad but was provided by the user. Details will be reported elsewhere. While motif discovery has been conditioned by a starting model that represents a family of related transcription factors, e.g. [4], in the work described here we used a PWM for the protein targeted by the ChIP-seq antibody, or for a protein that is known to interact with the target protein when the target associates with DNA. Given that our genome-wide ChIP-seq datasets had high spatial resolution and specificity, we anticipated that a seeded discovery method might a) identify variants on the dominant expected motif; b) distinguish motifs that were similar but biologically distinct within a single dataset, such as PDX1 and PDX1:PBX1 in islets data; and c) address issues related to the PDX1 motif being short and having a low apparent prevalence.

Because the seeded analysis involves only one starting position, it is computationally efficient. Given this, we set the number of expectation-maximization (EM) iterations to a relatively large number (80) for all runs. In the work described here, sequence centers corresponded to locations enrichment maxima in a ChIP-seq enrichment profile, and the spatial distribution of an expected motif should have a higher density near such enrichment maxima. To take advantage of this in motif discovery, we modified the GADEM's EM algorithm by weighting the likelihood with a triangular distribution whose maximum frequency was at sequence centers. In this modified version, for each discovery run, GADEM first varies the expected number of sites (referred to as *MAXP* in [5]) from 0.05 to 1.0 in increments of 0.05 (20 values in total) times the number of

sequences in the data. From the 20 resulting motifs (one for each *MAXP* value), it reports the one with the lowest E-value. It then automatically masks this motif's sites in the data and repeats the process until it can find no more motifs with an E-value below the threshold. Since the modified version used no spaced dyads, the genetic algorithm was unnecessary; the approach is equivalent to running the published version of GADEM with the genetic algorithm's number of 'generations' set to 1 and 'population' size to 20, with the user-specified starting PWM and with a range of *MAXP* values. For some datasets, we found that it was necessary to set a large, non-significant E-value threshold value (e.g., 10000), although the observed E-values for all motifs related to the target PWM were much lower. For example, the E-values for all variants of the FOXA2 and HNF4A motifs were highly significant (e.g. ln(E-value)=-9600).

From the set of motifs returned from a seeded discovery run, we reported the subset of motifs that were similar to the seed motif or to related PWMs (e.g. for known cofactors or protein complexes), typically using threshold value of 1.e-5 for a Pearson PWM similarity E-value [6]. When such subsets contained several motifs that appeared to be variants that could be represented as a general, global motif, we combined the variants into a single, general motif, using a custom C program that combined overlapping motif sites and ensured that each unique site was represented only once in the final combined motif. In combining motif variants, we used only the subset of variants that had the highest central densities in spatial distributions, in order to report a motif representing a high confidence subset of binding sites for a transcription factor. This being said, the other identified motifs we report but don't use for the merged motif are likely

biologically meaningful, but decreased specificity of the generated merged motif. In on going work we are addressing issues related to interpreting in more detail the range of sites identified for sets of motif variants.

To identify FOXA2 binding sites in mouse adult islet and liver data we used as a seed the 10-mer Foxa PWM as reported in Wederell et al 2008 [3]. We set GADEM's PWM score p-value limit to 2e-4 and ran the EM for 80 iterations or until convergence. For the islet data we retained the four motifs whose similarity E-values to the seed were less than 1e-8. Of these, we report results only for the motif (m2), which had the lowest discovery E-value, the largest number of sites, and the spatial distribution with the highest central density. This motif was found in 71% of the FOXA2 islet sites (Figure S5). For the liver data set we retained four motifs whose similarity E-values to the seed were less than ~1e-5, and merged sites for the two motifs (m2 and m5) that had the lowest discovery E-values and spatial distributions with the highest central densities. The final motif was found in 75% of the FOXA2 liver sites (Figure S6).

To identify PDX1 and PBX1 binding sites in MM0388-islets data we used two different seed PWMs: IPF1_Q4_01, TRANSFAC M01013 one for PDX1 and PBX1_02, TRANSFAC M00124 for PBX1 (Matys 2006). We set GADEM's PWM score p-value limit to 5e-4 and ran the EM for 80 iterations or until convergence. We retained two PDX1-like motif from the run seeded with IPF1_Q4_1 and one motif from the run seeded with PBX1_02 (Figure S7). We report results for a) the PDX1-like motif (M01013-seeded m2, E-value 4.1e-3 to TRANSFAC IPF1 M00436), which was found in 44.8% of

PDX1 sites, and b) the PBX1-like motif (M00124-seeded m2, E-value 7.6e-5 to TRANSFAC PBX-1b M01017), which was found in 41.9% of PDX1 sites. The sequence logo of the PBX1-like motif was consistent with the consensus sequence for a PDX1:PBX1 dimer (Li JBC 08). Together, one or both of these motifs was found in 61.6% of PDX1 sites (Figure S7).

Finally, to identify HNF4A binding sites in adult liver we used TRANSFAC v9.3 HNF4_Q6_01 (M01031) as a seed PWM, because it was complied from a large number of known functional sites. As for FOXA2, we set GADEM's PWM score p-value threshold to 2e-4 and ran 80 iterations of EM or until convergence. We retained four HNF4A-like motif variants. The most frequent variant had a similarity E-value of 0.0 for M01031, and the other three variants had E-values of 4.5e-10, 3.1e-5 and 9.0e-4. The merged motif from combining all four variants had an E-value of 0.0 for M01031, and was found in 91.6% of the HNF4A sequences (Figure S8).

Quantitative real-time PCR

Primers were designed using Primer3. Primer sequences are available upon request. For ChIP-qPCR, DNA from triplicate ChIP experiments was obtained and amplified using an ABI 7500 PCR system (Applied Biosystems) and SYBR® Green supermix (Applied Biosystems). The fold enrichment of each target site was calculated as $2^{\Delta cT}$ between rabbit IgG and anti-FOXA2, anti-PDX1, or anti-HNF4A immunoprecipitated samples.

Supplementary Table

Table S1: Summary of the ChIP-seq libraries used in this study

		Total reads sequenced (M)	Mapping %	Reads Mapped (M)	Threshold	Reads in enriched regions (M)	# of regions	# of filtered[1] regions
FOXA2	Islet	40.2	38	15.2	9	0.19	7,409	7,189
FOXA2	Liver	34.4	33	11.4	9	0.21	10,970	10,701
PDX1	Islet	62.1	24	14.9	11	0.41	13,711	13,448
HNF4A	Liver	24	53	12.8	14	0.39	12,833	12,494
H3K4me1	Islet	19.3	66	12.7				
H3K4me1	Liver	37.9	71	27.3				
H3K4me3	Islet	8.4	55	4.6				
H3K4me3	Liver	6.5	54	3.5				

[1]Filtered refers to the removal of regions overlapping enriched sites identified in using an input control library

Supplementary Figures

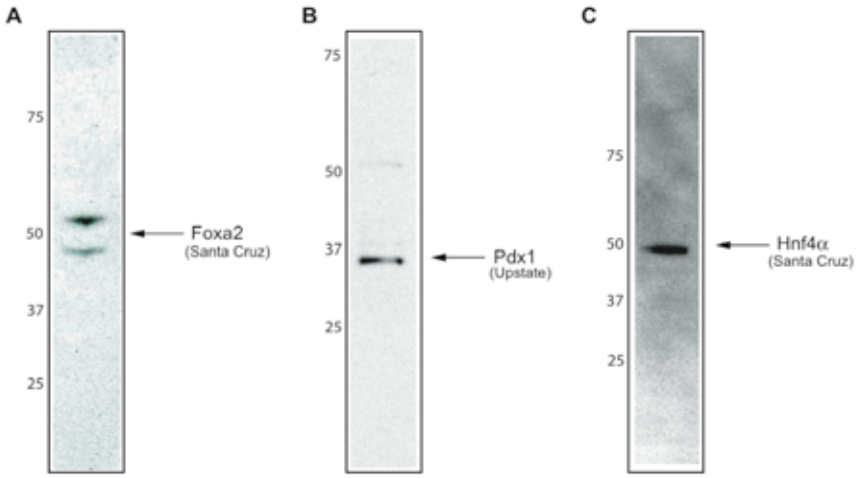


Figure S1: Specificity of Antibodies Used for ChIP-Seq Libraries. MIN6 cell lysates were used to perform Western Blots with the antibodies of interest: (A) FOXA2 (Santa Cruz sc-6554), (B) PDX1 (Upstate 07-696), and (C) HNF4A (Santa Cruz sc-8987). Clean bands in the expected size ranges were observed for each of the antibodies.

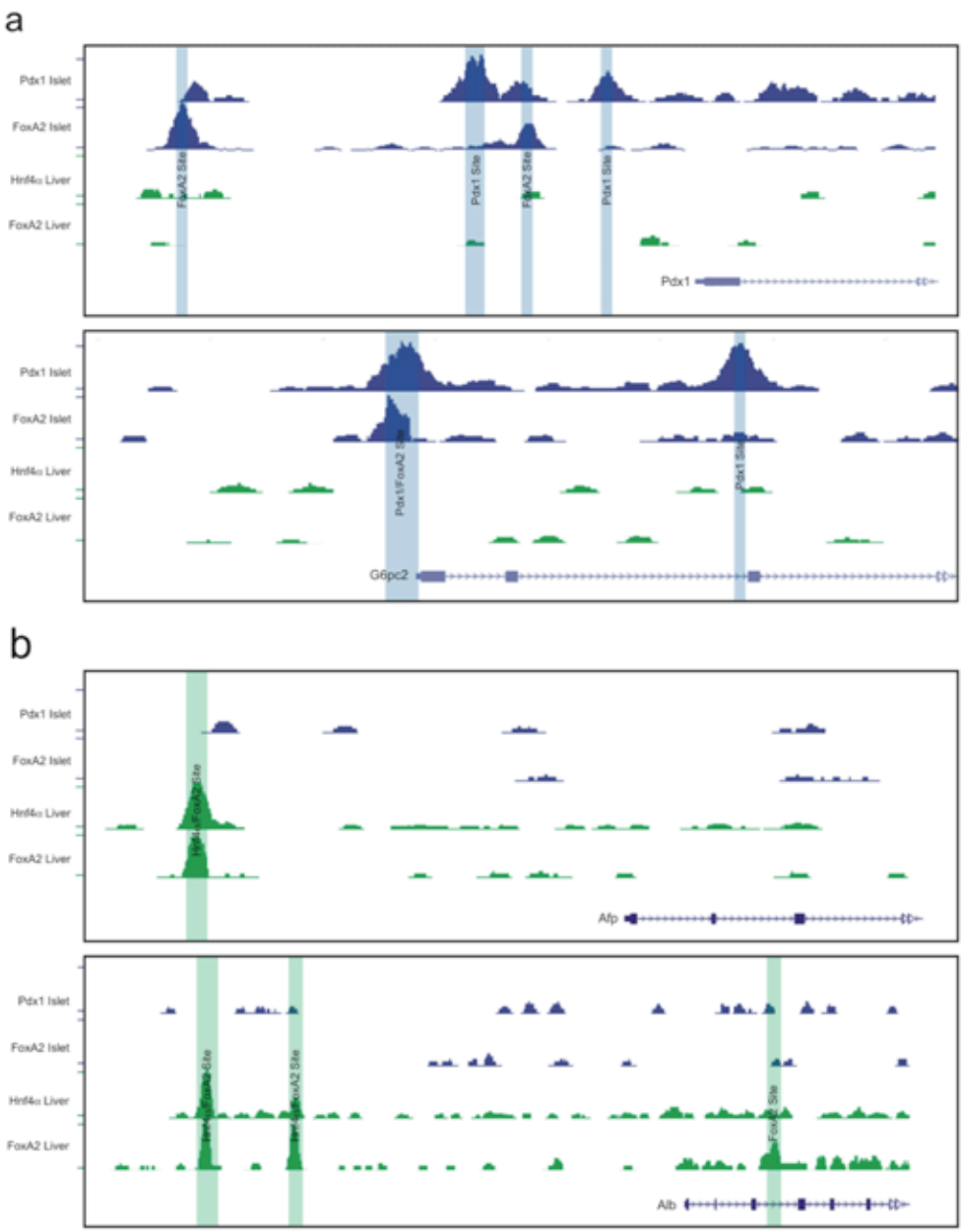


Figure S2: Identification of FOXA2, PDX1 and HNF4A occupied loci. UCSC mm8 genome browser screenshots of regions containing (A) FOXA2 and PDX1 sites in pancreas islets (blue) or (B) FOXA2 and HNF4A sites in liver (green).

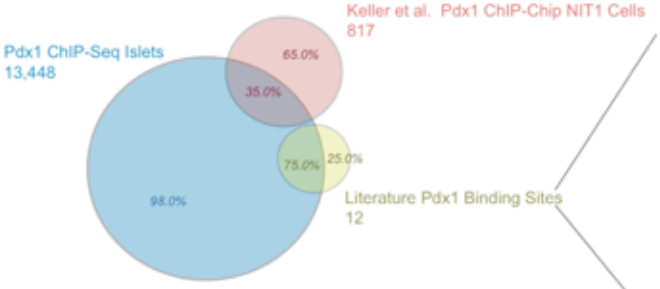
A

Foxa2 ChIP-Seq Peaks are Present at Known Foxa2 Binding Sites

Gene	Site Co-ordinates	Islet or Cell Line	Reference	ChIP Peak
Mafa	-7943 to -7910	Btc3 cell line	Raum JC, et al. 2006	No
Pdx1	-2153 to -1923	INS1 cell line	Samares SE, et al. 2002	Yes
Hadhsc	Intron 1	Islets	Lantz KA, et al. 2004	Yes
Gcg	G1/G2 Promoter element	InR1G9 cell line	Gauthier BR, et al. 2002	No*
Glut2	+87 to +132	Cell lines	Cha JY, et al. 2000	No
Kir6.2	-1364 to -1210	IEC-6, RIN-5F cell lines	Hashimoto T, et al. 2005	No**
Nkx2.2	Exon1a promoter	Btc3 cell line	Watada H, et al. 2003	Yes

* The absence of a Foxa2 binding peak at Gcg is not unexpected as Gcg is expressed in alpha cells which compose only 10-15% of islet mass.
** ChIP-qPCR of the Kir6.2 target produced strong enrichment values, indicating that it is clearly valid and missed by our ChIP-Seq analysis. We do observe a peak at the Foxa2 binding site for Kir6.2, however its height was below our thresholded level. It is likely that deeper sequencing would reveal a Foxa2 peak.

B



Gene	Reference	ChIP Peak
Ins1 A1	German et al. 1995	Yes
Ins1 A3/4	German et al. 1995	Yes
Ins2 A1	German et al. 1995	Yes
Ins2 A3/4	German et al. 1995	Yes
Pdx1	Gerrish et al. 2001	Yes
Gck	Shelton et al. 1992	Yes
Glut2	Waeber et al. 1996	No*
IAPP A1	Carty et al. 1997	Yes
IAPP A2	Carty et al. 1997	Yes
Sst TSE1	Leonard et al. 1993	No**
Sst TSE2	Leonard et al. 1993	No**
Mafa	Raum et al. 2006	Yes

* We do identify Pdx1 binding peaks at Glut2 but not at the exact location outlined by Waeber.
** The absence of Pdx1 binding peaks at Sst in our data is not unexpected as Sst is expressed in delta cells which compose a minimal % of islet mass.

C

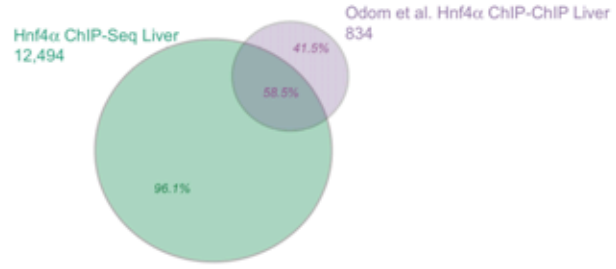


Figure S3: ChIP-Seq Libraries Correlate to Known Binding Sites. (A) A table showing a literature survey of known FOXA2 binding sites in islets, and whether we observe a FOXA2 enriched regions at the site [7-13]. The table reveals that our data identifies enriched regions at most of the known sites. (B) PDX1 islet enriched regions were compared against PDX1 ChIP-Chip binding tiles identified from a NIT1 cell line [14].

To determine similar regions between the two, we looked for overlap between ChIP-Chip binding tiles and 500 base pairs regions flanking points of maximal enrichment. Using this method, 35% of ChIP-Chip binding regions were identified in our data. Although this level of overlap is slightly less than previously reported when comparing ChIP-Seq to ChIP-Chip a further comparison with 12 known PDX1 binding sites reveals that 9 are present in our data set, while none are represented in the ChIP-Chip regions [11, 15-20]. Furthermore, of the 3 known sites not identified by in our data, 2 of them are associated with somatostatin which is expressed in delta cells, while the other (Glut2) does have an associated PDX1 site but not at the exact region suggested by literature. (C) HNF4A liver ChIP-Seq binding peaks were compared against HNF4A ChIP-Chip results in the same manner as described above [22]. An overlap of 58.5% of the ChIP-Chip regions was observed, a percentage consistent with previously reported comparisons of ChIP-Seq and ChIP-Chip [3]. It should be noted that we do not detect all of the regions found in ChIP-chip studies. This is a result of a lack of saturation in our analyses (Supplemental Fig. 16) and as the arrays used in ChIP-chip studies only represent a fraction of the genome they are often able to detect low occupancy sites not picked up in ChIP-seq studies.

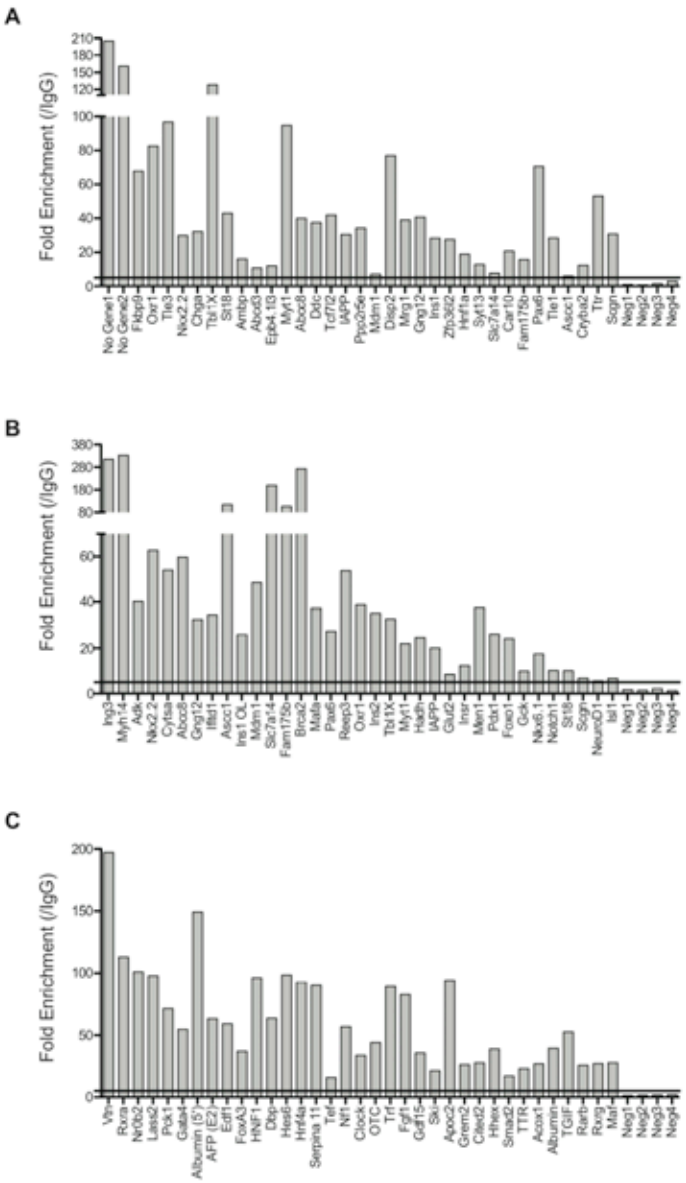


Figure S4: ChIP-qPCR Validations of Libraries. ChIP-qPCR was performed for (A) FOXA2 in islets, (B) PDX1 in islets, and (C) HNF4A in liver to validate targets identified by our ChIP-Seq data. 35 targets were tested for each antibody and in every case all 35 showed enrichment over 4 negative controls compared to IgG. Primer sequences are available upon request.

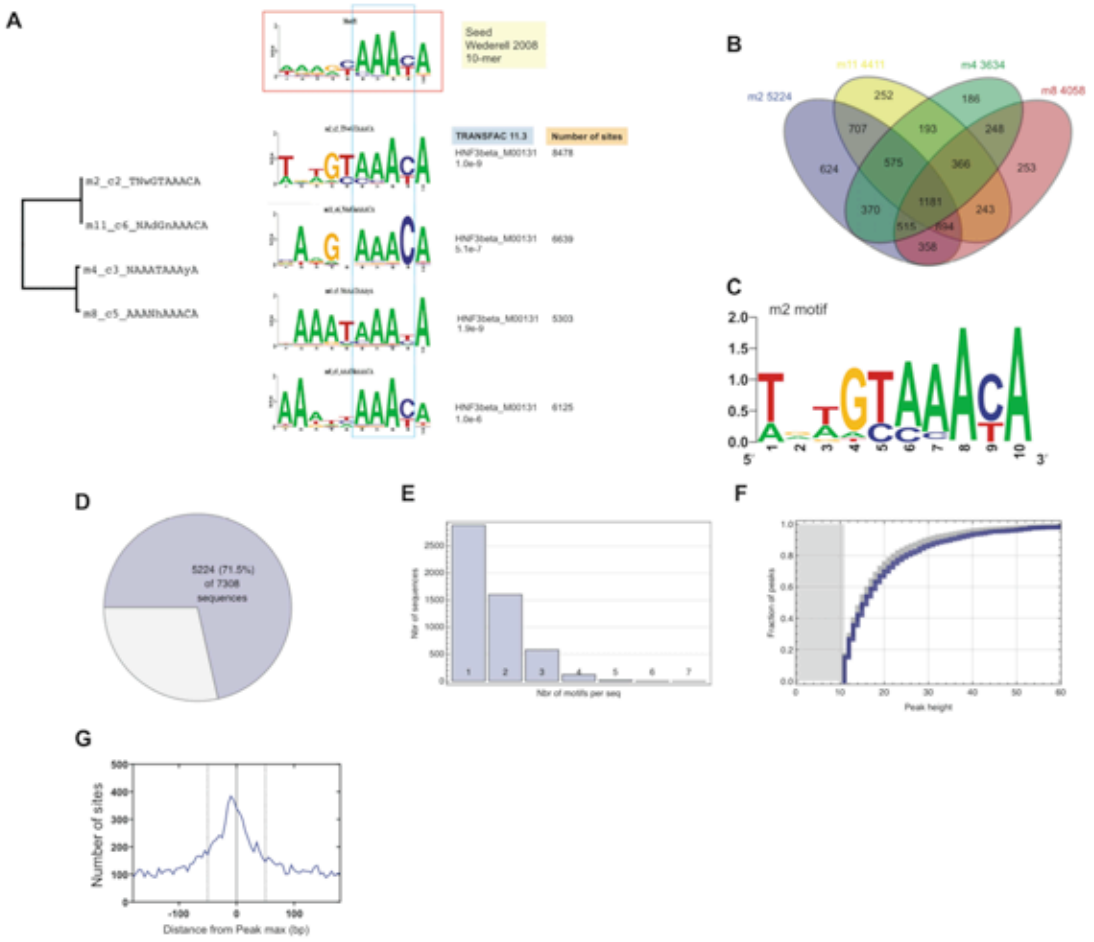


Figure S5: Seeded motif discovery on Identified FOXA2 sites in Islets. The 10-mer motif from Wederell et al was used as a seed [3] for motif discovery with GADEM [1] on sequences extracted from 400-bp regions that were centered on locations of maximal FOXA2 enrichment in islets. (A) Four motif variants were returned that showed high similarity to the seed based on STAMP E-values. The figure shows a tree of PWM similarities to each other, gives E-values to relevant TRANSFAC motifs, and the number of sites for each motif. (B) Venn diagram of the distribution of motif variants across sequences. (C) The 10-bp sequence logo for motif variant m2. Because the spatial distribution of m2 sites had the highest density near sequence centers (data not shown),

its sites represent a high-confidence subset of potential FOXA2 binding sites. (D) The fraction of sequences that have at least one m2 motif site. (E) The distribution of the number of m2 sites in 400-bp sequences, for sequences with at least one such site. (F) Cumulative distributions of the fraction of enriched regions as a function of region score (peak height) at least one m2 site (blue line), compared to the overall set of FDR-thresholded input sequences (grey line). (G) Distributions of the motifs site sequences around peak maxima.

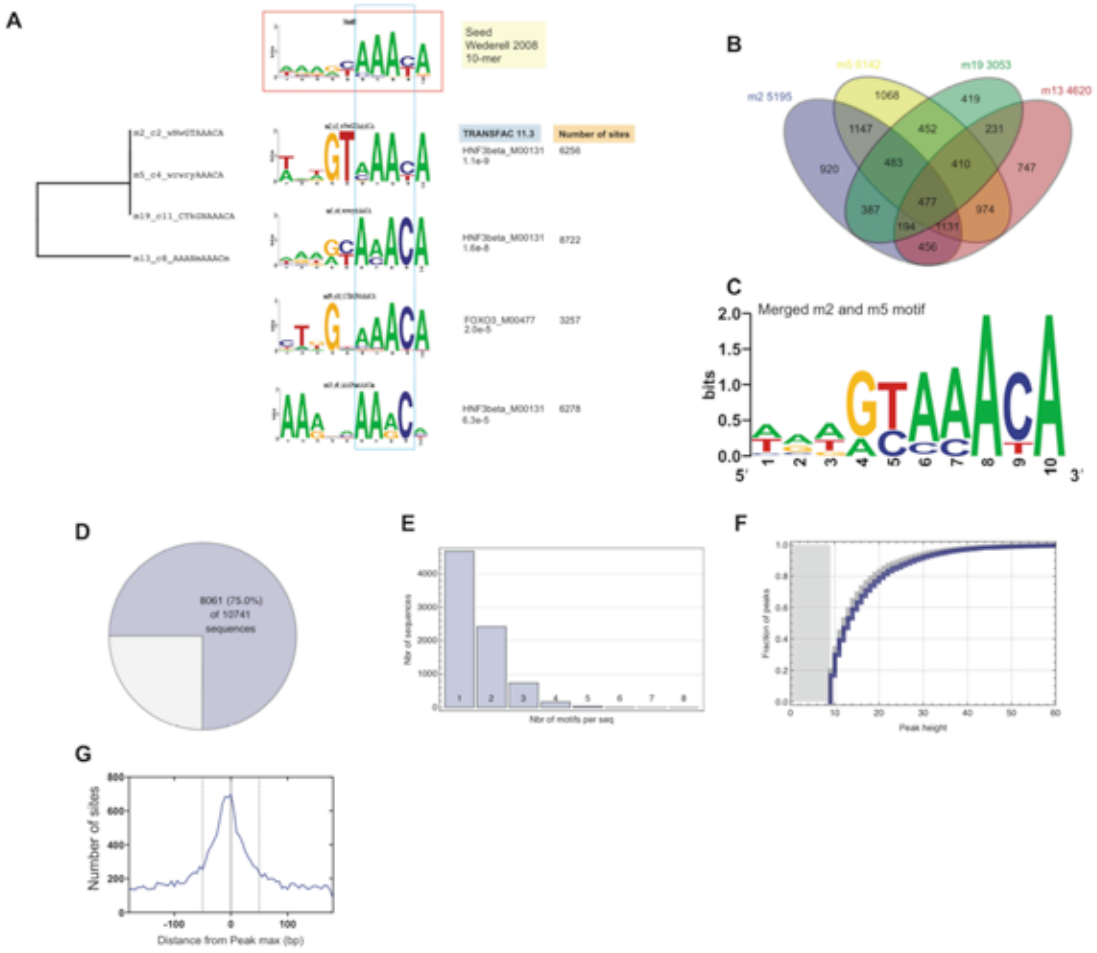


Figure S6: Seeded motif discovery on Identified FOXA2 sites in Liver. The 10-mer motif from Wederell et al was used as a seed [3] for motif discovery with GADEM [1] on sequences extracted from 400-bp regions that were centered on locations of maximal FOXA2 enrichment in liver. (A) Four motif variants were returned that showed high similarity to the seed based on STAMP E-values. The figure shows a tree of PWM similarities to each other, gives E-values to relevant TRANSFAC motifs, and the number of sites for each motif. (B) Venn diagram of the distribution of motif variants across sequences. (C) The 10-bp sequence logo for a motif generated by merging sites for variants m2 and m5. Because the spatial distribution of sites for these two variants had

the highest density near sequence centers (data not shown), the set of merged sites represent a high-confidence subset of potential FOXA2 binding sites. (D) The fraction of sequences that have at least one merged motif site. (E) The distribution of the number of merged motif sites in 400-bp sequences, for sequences with at least one such site. (F) Cumulative distributions of the fraction of enriched regions as a function of region score (peak height) at least one merged motif site (blue line), compared to the overall set of FDR-thresholded input sequences (grey line). (G) Distributions of the motifs site sequences around peak maxima.

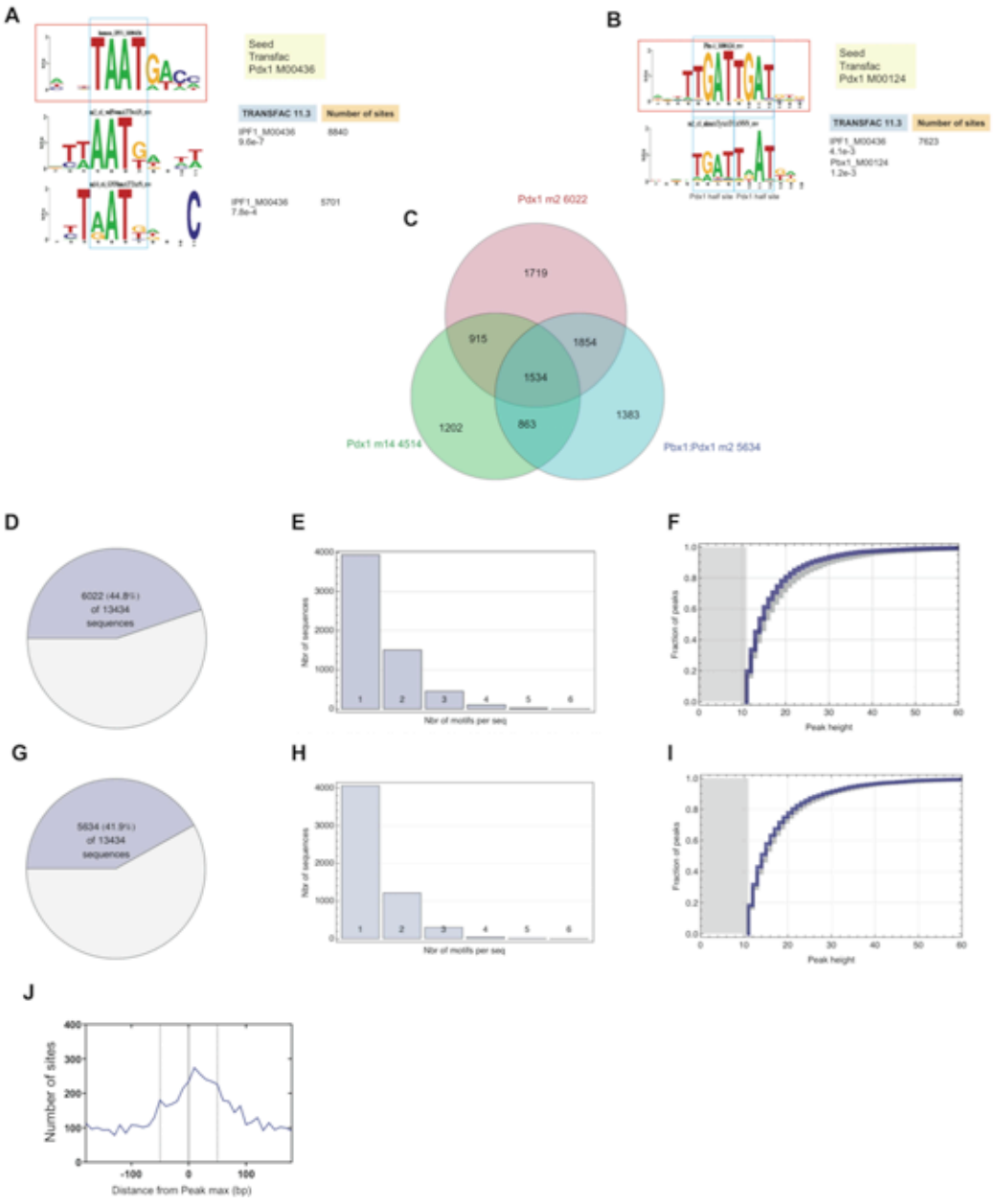


Figure S7: Seeded motif discovery on identified PDX1 sites in Islets. Both the 11-bp PDX1 M00436 motif and the 15-bp M00124 PBX1-1 from TRANSFAC were used as seeds for motif discovery with GADEM [1] runs on sequences extracted from 400-bp regions that were centered on locations of maximal enrichment of PDX1 in mouse adult islets. (A) The PDX1 seed returned two motifs that showed high similarity to the seed

based on STAMP E-values, and (B) the PBX1 seed returned one motif that showed high similarity to the seed based on STAMP E-values. These figures show the similarity E-values to relevant TRANSFAC motifs, and the number of sites for each motif. (C) Venn diagram of the distribution of the PBX1- and PDX1-like motifs across sequences. While the spatial distribution of sites for these two variants had the highest density near sequence centers, central densities were more modest than for FOXA2 in both tissues and for HNF4A in liver. Because it had a more centrally dense spatial distribution than m14, we used only the m2 PDX1-like motif to represent a set of potential PDX1 binding sites. (D) The fraction of sites with a PDX1 monomer-like m2 motif. (E) The number of sites with the indicated number of PDX1 monomer-like m2 motifs. (F) Cumulative distribution plot of the fraction of peaks with at least one PDX1 monomer-like m2 motif (blue line) as compared to the overall set of sequences (grey line), as a function of peak height. (G) The fraction of sites with a PDX1:PBX1 dimer-like motif. (H) The number of sites with the indicated number of PDX1:PBX1 dimer-like motifs. (I) Cumulative distribution plot of the fraction of peaks with at least one PDX1:PBX1 dimer-like motif (blue line) as compared to the overall set of sequences (grey line), as a function of peak height. (J) Distributions of the motifs site sequences around peak maxima.

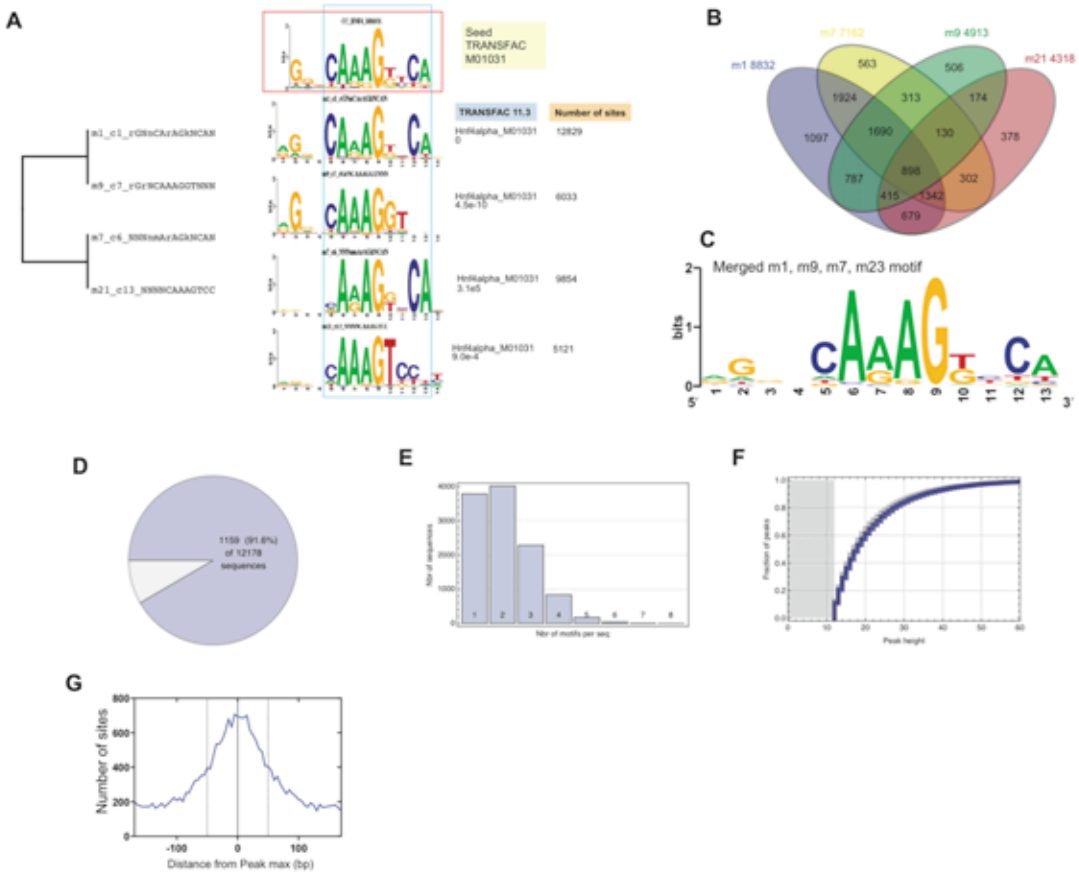


Figure S8: Seeded De novo motif discovery on Identified HNF4A sites in Liver. The 13-bp motif from TRANSFAC v9.3 M01031was used as a seed for motif discovery with GADEM [1] runs on sequences extracted from 400-bp regions that were centered on locations of maximal HNF4A enrichment in mouse adult liver. (A) Five motif variants were returned that showed high similarity to the seed based on STAMP E-values, except for having little information for three to four positions at either their 5' or 3' ends. The figure shows a tree of PWM similarities to each other, gives E-values to relevant TRANSFAC motifs, and the number of sites for each motif. (B) Venn diagram of motif variants on sequences. (C) The 13-bp sequence logo for a motif generated by merging sites for all five variants. Because for all variants the spatial distribution of sites had high

densities near sequence centers (data not shown), the overall set of merged sites represents a high-confidence set of potential HNF4A binding sites. (D) The fraction of sequences that had at least one merged motif site. (E) The distribution of the number of merged motif sites in 400-bp sequences, for sequences with at least one such site. (F) Cumulative distributions of the fraction of peaks with at least one merged motif site (blue line) as compared to the overall set of sequences (grey line), as a function of peak height threshold. (G) Distributions of the motifs site sequences around peak maxima.

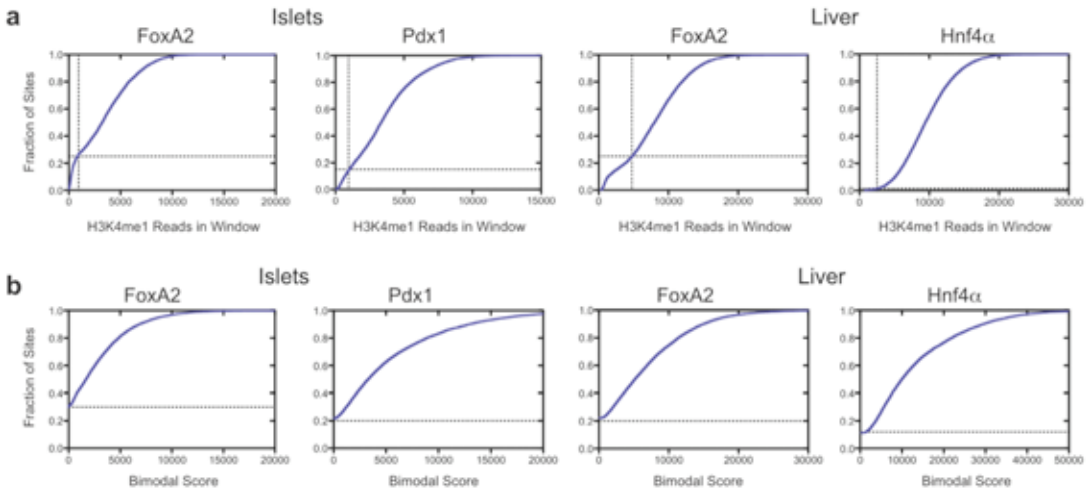


Figure S9: Discrimination of H3K4me1 site classes. (A) A cumulative distribution plot of the total number of H3K4me1 sequence reads in a +/-2 Kb window flanking identified peak maxima. The indicated inflection points (dotted lines) in the curves were used to determine minimum threshold values for a site to be considered to be associated with H3K4me1. Sites below this threshold were considered to be in the low H3K4me1 class. For the islet data this threshold value was 950 reads for the FOXA2 and PDX1 libraries, which called 1954 and1803 sites in the low H3K4me1 class, respectively. For the liver data the threshold was 4750 for the FOXA2 liver library, and 2500 for the HNF4A library, calling 2271 and 169 peaks in the low H3K4me1 class, respectively. (B) A cumulative distribution plot of the area of the virtual triangle drawn between the flanking H3K4me1 enrichment maxima and the central trough minimum, or bimodal score. For this only sites that were above the total H3K4me1 read threshold were considered. Peaks with bimodal score of 0 were considered monomodal. The islet FOXA2 and PDX1 libraries contained 1250 and 2300 monomodal sites, while the liver FOXA2 and Hnfα libraries contained 1337 and 1465 monomodal sites.

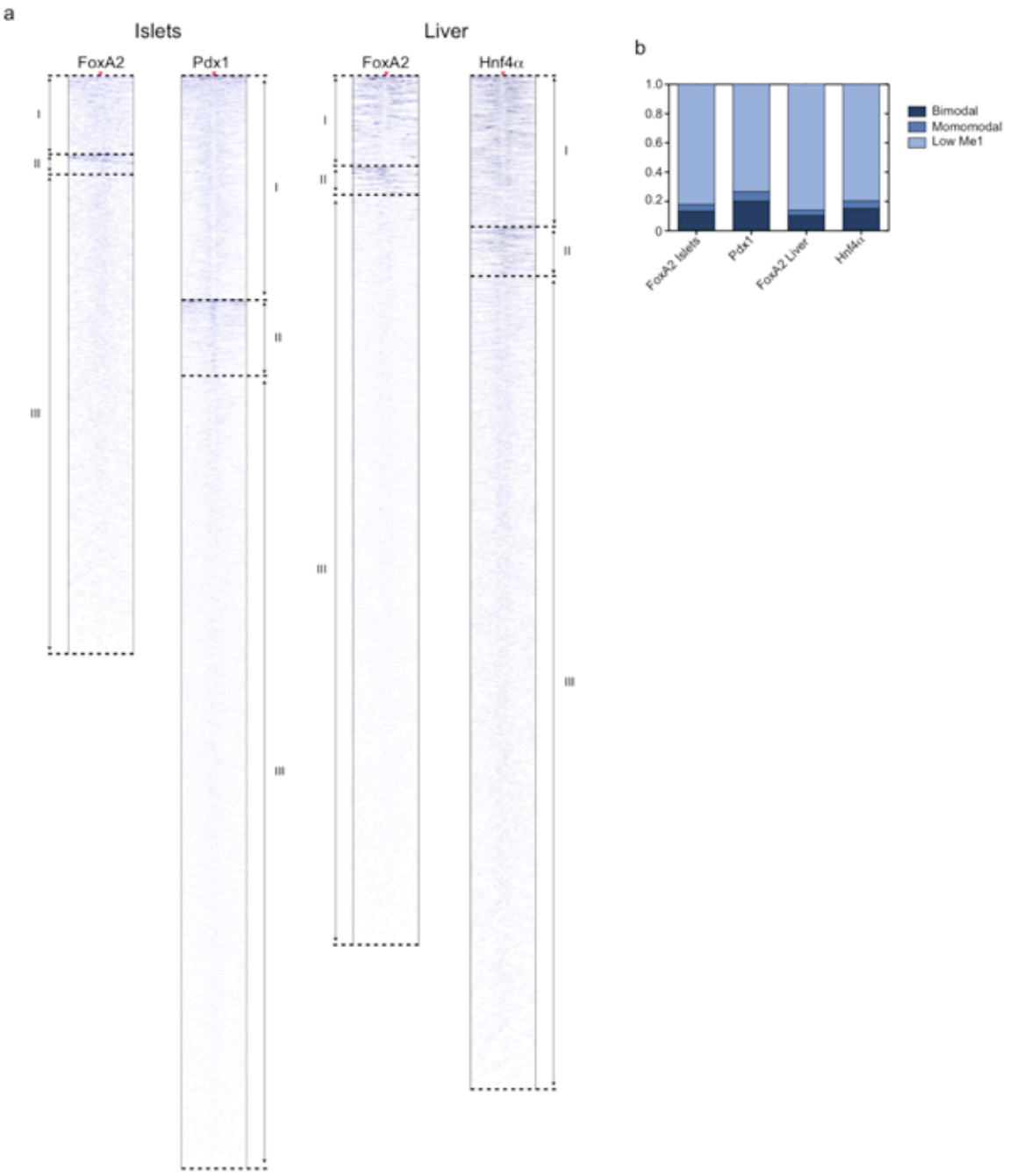


Figure S10: H3K4me3 profiles fail to globally discriminate transcription factor site classes. (A) Heatmaps of H3K4me3 read density in ±2-kb regions centered on FOXA2 (islets or liver), PDX1 and HNF4A peak maxima. Peak max locations are indicated by red triangles, with flanking H3K4me1 read density plotted horizontally in blue for each peak. H3K4me1 read density is represented by the intensity of blue in the heatmaps: dark

blue indicates high, and light blue indicates low read density, while white indicates minimal or no H3K4me1. The grouping of sites into H3K4me3 classes is indicated, with class I indicating bimodal sites, class II indicating monomodal sites, and class III indicating low H3K4me3 sites (B) Fractional representation of the population of bimodal, monomodal, or low H3K4me3 sites peaks in each library. Significantly, H3K4me3 profiling calls 83% (FOXA2 Islets), 73% (PDX1), 85% (FOXA2 Liver), and 80% (HNF4A) enriched regions in the low H3K4me3 class.

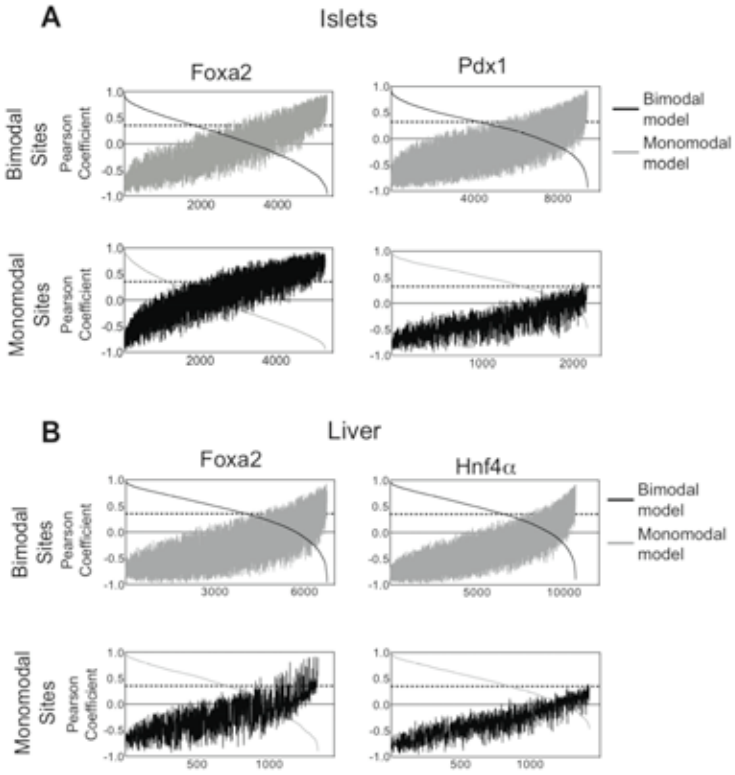


Figure S11: Determination of the high confidence H3K4me1 enrichment class call peak subset. Plots of ranked pearson correlation coefficients obtained from comparisons of H3K4me1 profiles in the indicated classes (left) with the mean bimodal and monomodal profile models in (A) islets and (B) liver. Based on these plots a correlation threshold was chosen (dotted lines) to identify sites with a high correlation to the correct model and a low correlation with the alternative profile model.

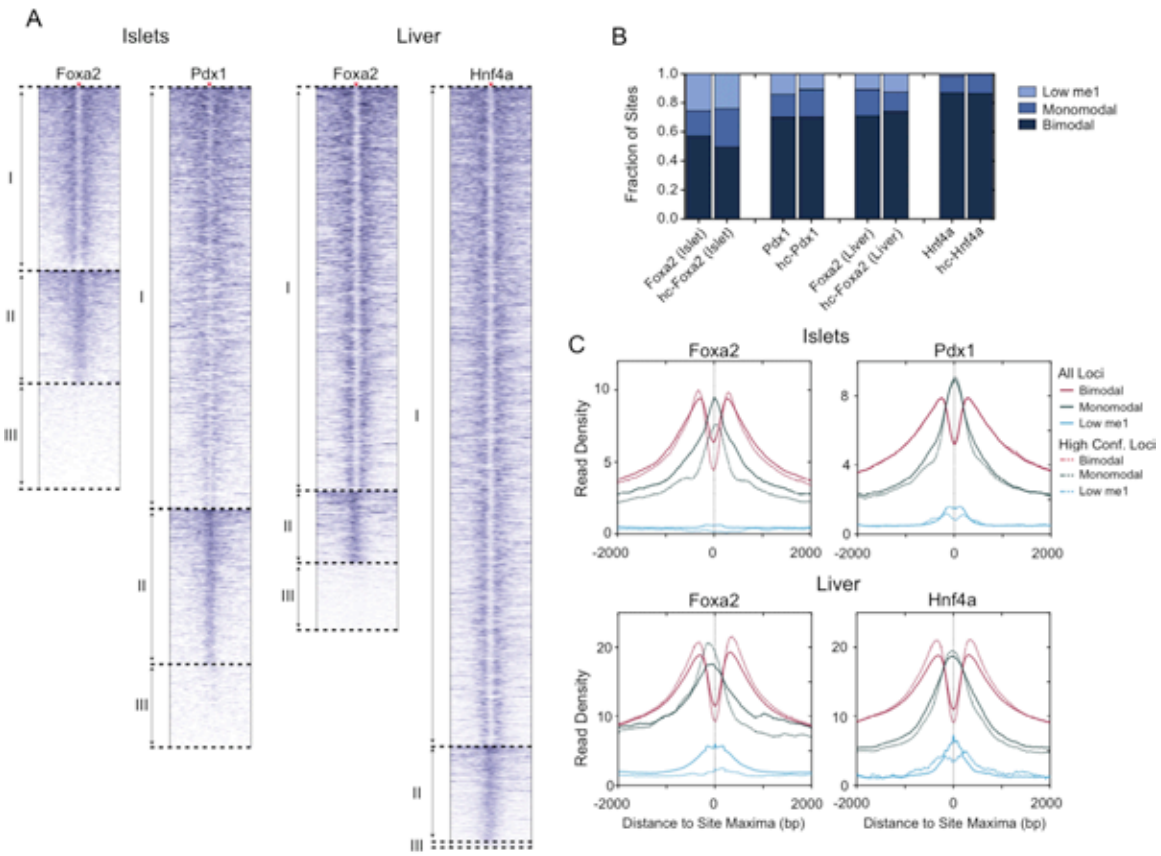


Figure S12. High Confidence H3K4me1 profile classes have similar profiles and distributions as H3K4me1 profile classes based on all loci. (A) Heatmaps showing H3K4me1 profiles of high confidence sites, with class I indicating bimodal sites, class II indicating monomodal sites, and class III indicating low H3K4me3 sites. Heatmaps are of H3K4me1 read density in ±2-kb regions centered on high confidence peak maxima. Peak max locations are indicated by red triangles, with flanking H3K4me1 read density plotted horizontally in blue for each site. H3K4me1 read density is represented by the intensity of blue in the heatmaps: dark blue indicates high, and light blue indicates low read density, while white indicates minimal or no H3K4me1. (B) Fractions of bimodal, monomodal, or low H3K4me1 sites in each library, for all sites, and for the high

confidence (-hc) sites. (C) Mean H3K4me1 enrichment profiles associated with each site class in the FOXA2 (islets or liver), PDX1, and HNF4A peak sets. Peak maxima are centered at 0. Solid lines show average profiles for all peaks in a class, while dotted lines show average profiles for high confidence peaks.

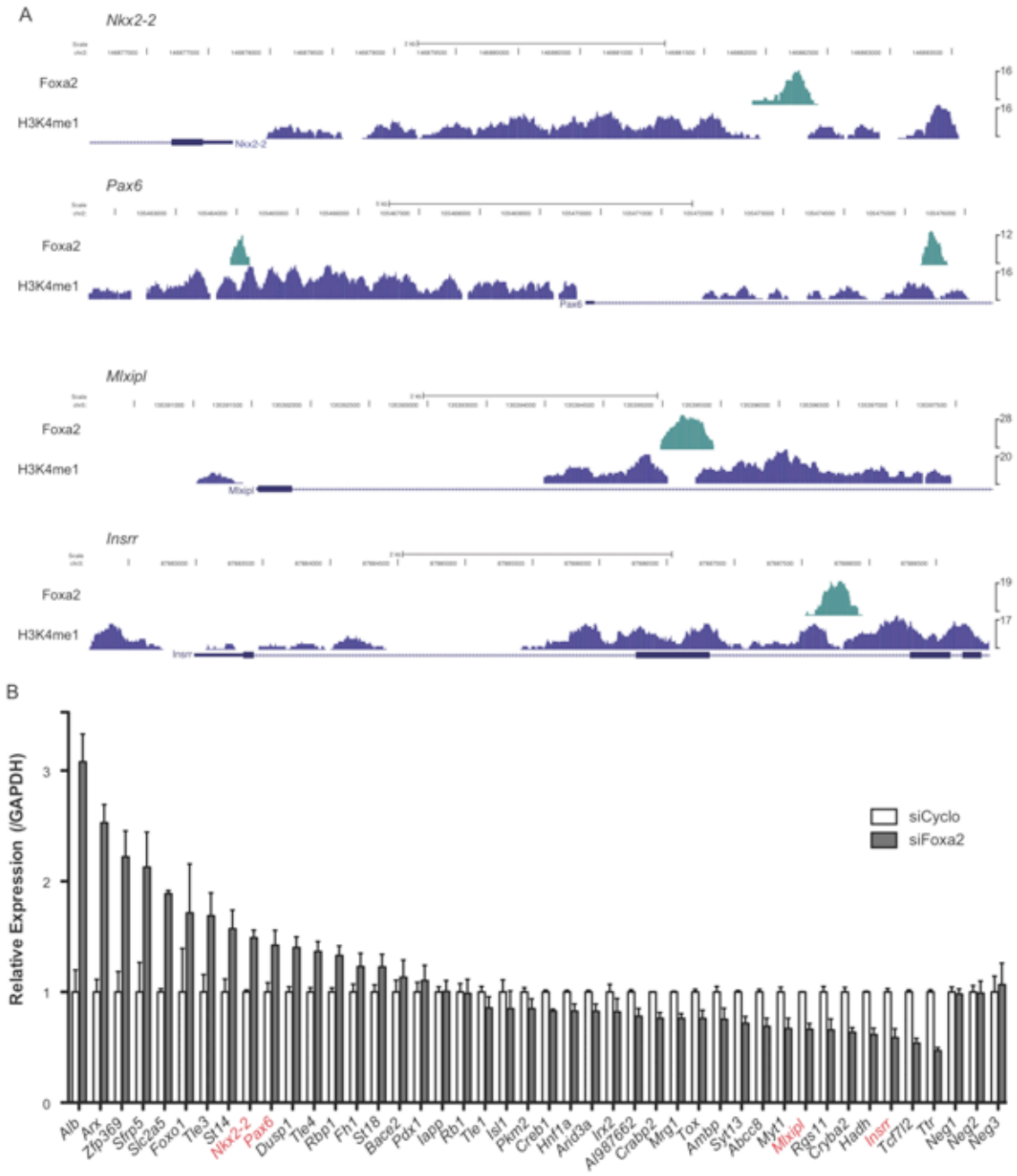


Figure S13: Bimodal FOXA2 sites in islets are associated with genes that are islet specific and that are regulated by FOXA2. (A) UCSC mm8 genome browser screenshots of regions containing bimodal FOXA2 sites near genes that are highly expressed and highly specific to islets. (B) *Foxa2* suppression alters the expression of

genes with bimodal sites. The relative expression levels of the indicated genes as detected by qRT-PCR in islets treated with siRNAs targeting *Foxa2* as compared to islets treated with the *siCONTROL Cyclophilin B* siRNA reagent. The expression levels of three genes without an associated *Foxa2* peak are also shown. The names of the genes shown in (A) are highlighted in red.

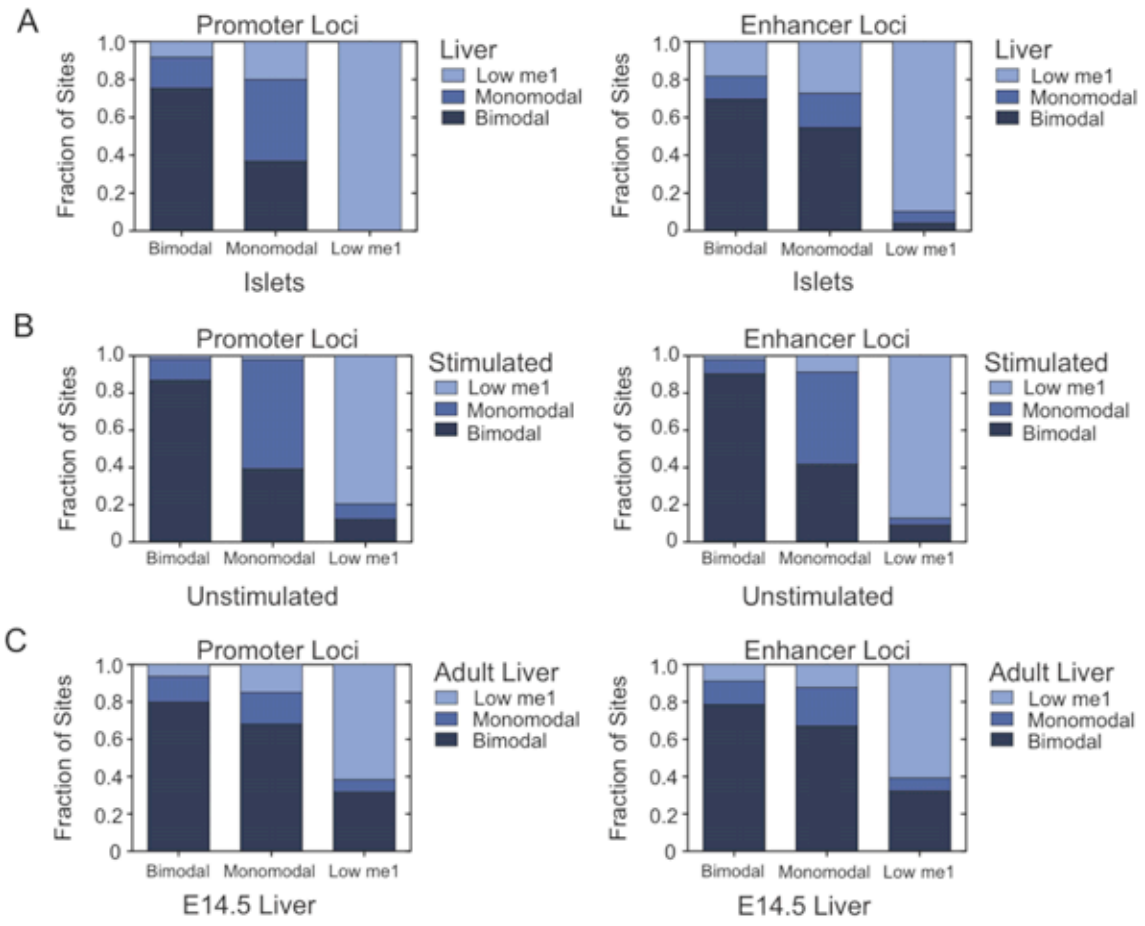


Figure S14: The occupancy of loci by H3K4me1 marked nucleosomes in both promoters and enhancers is altered by lineage, signaling events and development. (A) The fraction of loci bound by FOXA2 in islets and liver in promoters or enhancers that are bimodal, monomodal, or low H3K4me1 in islet versus liver. (B) The fraction of loci bound by STAT1 in IFNG-stimulated and unstimulated HeLa cells in promoters or enhancers that are bimodal, monomodal, or low H3K4me1 in IFNG-stimulated versus unstimulated HeLa. (C) The fraction of loci bound by FOXA2 in e14.5 liver and adult liver in promoters or enhancers that are bimodal, monomodal, or low H3K4me1 in e14.5 liver and adult liver.

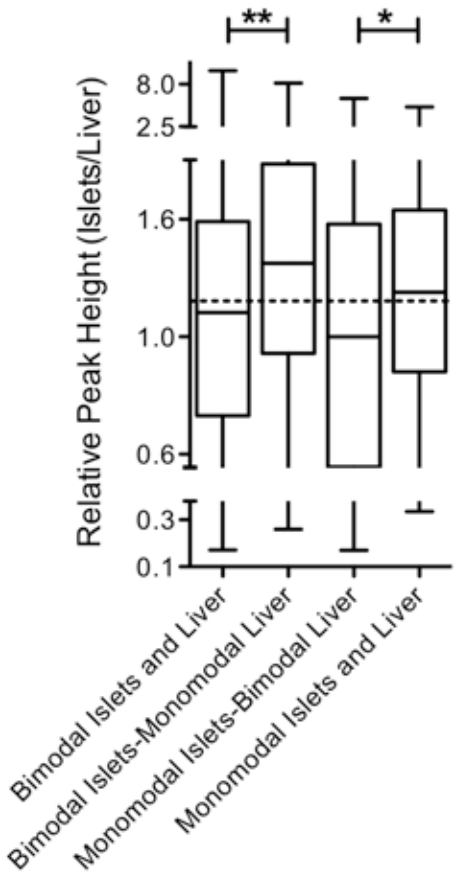


Figure S15: Shared FOXA2 binding sites bimodal in one tissue and monomodal in the other are more occupied in the tissue with the bimodal site. Box-whisker plot of the relative peak heights (islets/liver) of FOXA2 sites found in both islets and liver. The dotted line indicates the median of relative peak height of all shared FOXA2 sites. Differences in peak height were assessed using a Kruskall Wallace non-parametric test with a Dunns comparison. * indicates $p<0.05$, while ** indicates $p<0.01$.

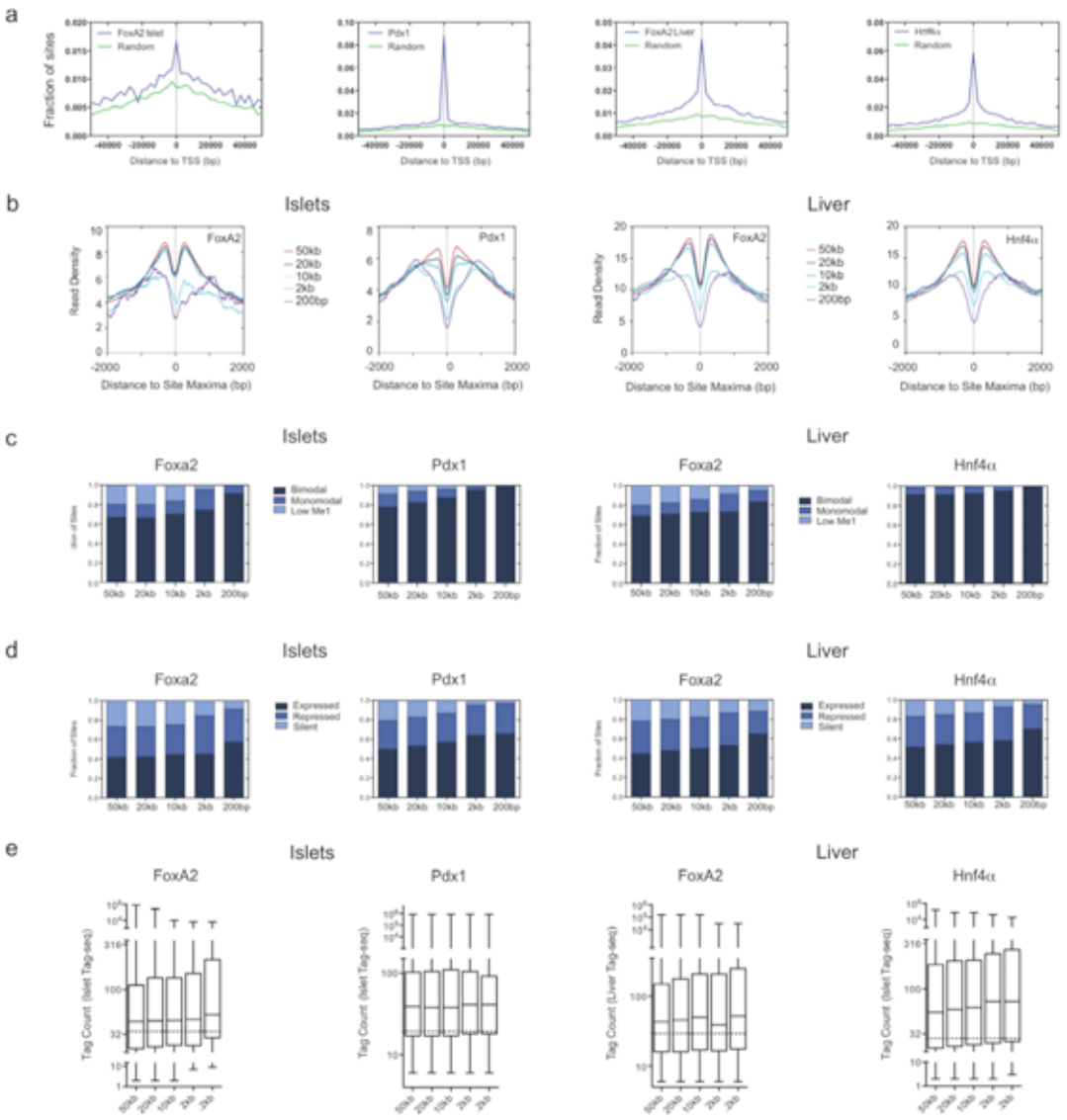


Figure S16: Effect of gene association distance on enriched region attributes. (A) Distribution of identified peak maxima around Refseq TSS's, regions +/- 50 Kb are shown. Note that four all four transcription factor data sets the distribution is above random. (B) Average H3K4me1 enrichment profiles of sites within indicated distances of a Refseq TSS. (C) Fraction of bimodal, monomodal, or low H3K4me1 enriched regions at the indicated distances. Fraction of enriched regions associated with expressed, not

expressed, or silent genes. (D) Box-whisker plots of tag counts of Islet or Liver expressed genes associated with a TF at each distance. No statistical difference in tag count was observed. These data indicate that sites in promoter regions are more commonly bimodal and associated with expressed genes. However, site-gene association distance has little affect when enhancer regions are considered.

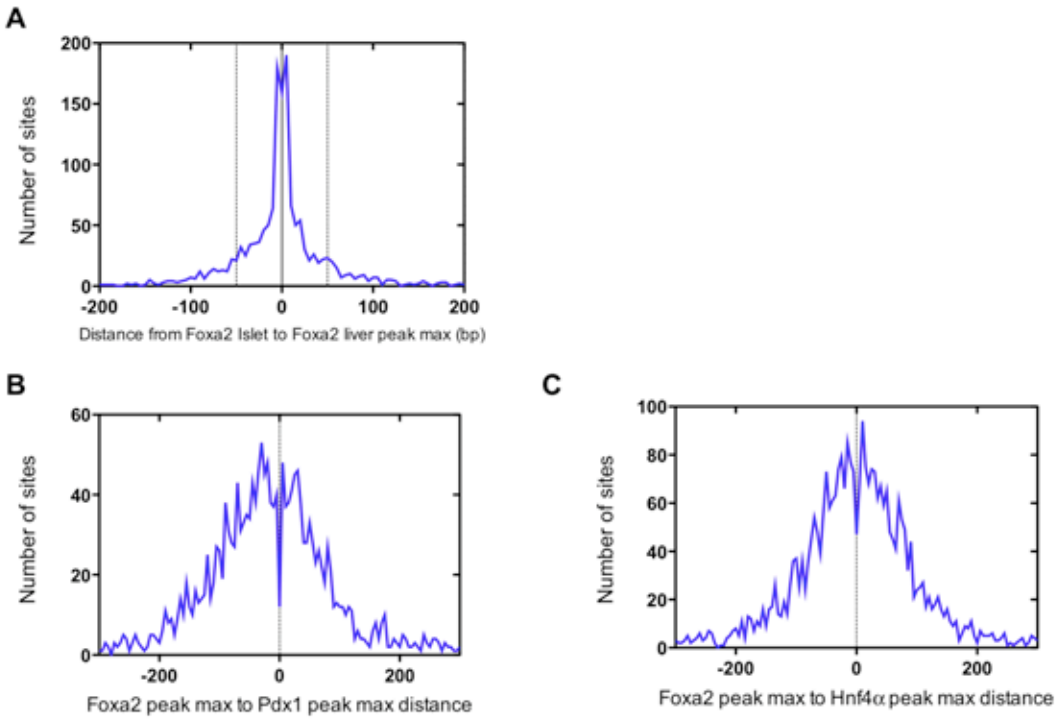


Figure S17: Distribution of binding sites around FOXA2 sites. (A) Distribution of FOXA2 liver sites around locations of maximal enrichment of FOXA2 islet sites. (B) Distribution of PDX1 sites around locations of maximal enrichment of FOXA2 islet sites. (C) Distribution of HNF4A sites around locations of maximal enrichment of FOXA2 liver sites.

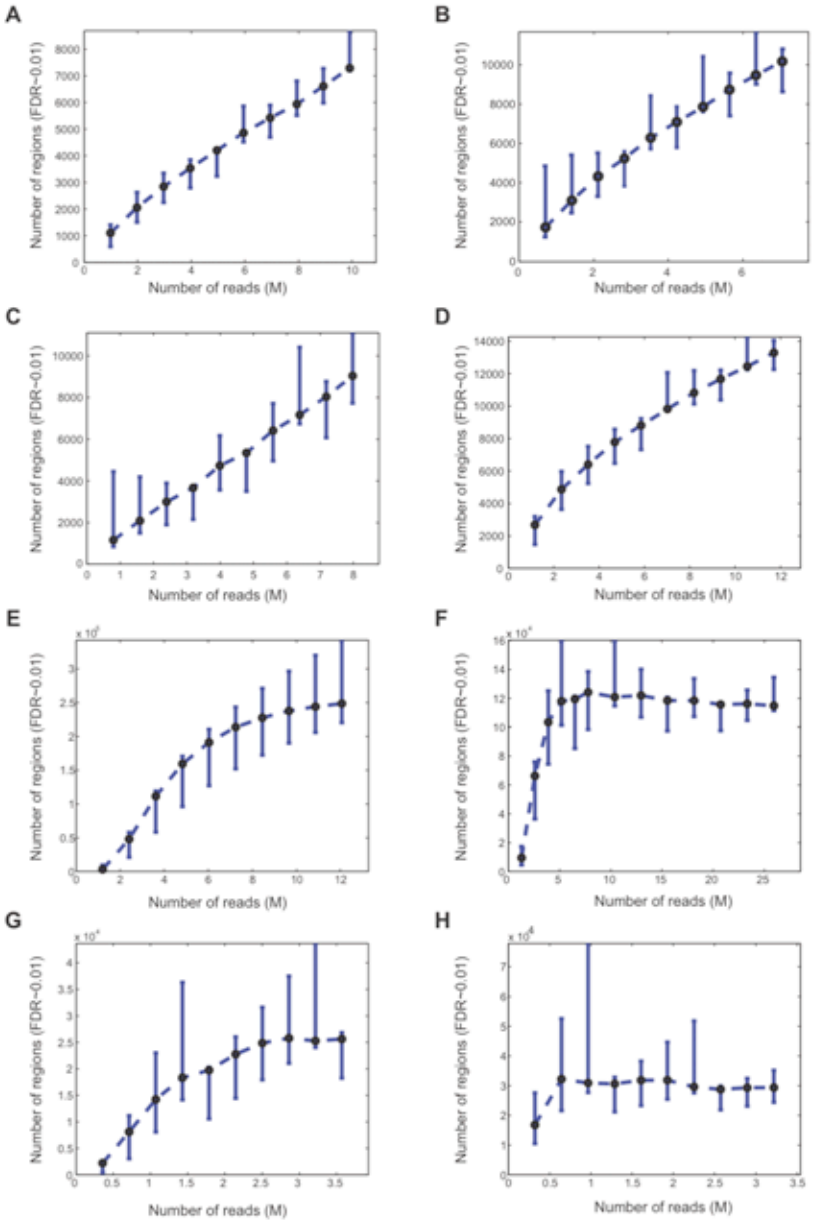


Figure S18: Transcription factor ChIP-seq libraries are not saturated but histone methyl mark libraries are. Saturation analysis of sequence reads from FOXA2 in (A) islets, (B) liver, (C) PDX1 in islets, (D) HNF4A in liver, H3K4me1 in (E) islets, (F) liver, or H3K4me3 in (G) islets, (H) liver. The numbers of enriched regions at FDR 0.01 is represented by circular symbols, and were linearly interpolated from the numbers

estimated for the two FDR values flanking FDR~0.01 which corresponded to integer-valued height thresholds. For each point we performed five randomizations. Results were very repeatable and variations for different subsamples fell within the circular plot symbols. Error bars showed average number of enriched regions predicted for two flanking FDR values.

Supplementary References

1. Li, L. GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *J Comput Biol* **16**, 317-329 (2009).
2. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**, D108-110 (2006).
3. Wederell, E.D. *et al.* Global analysis of in vivo FOXA2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res* **36**, 4549-4564 (2008).
4. Sandelin, A. & Wasserman, W.W. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol* **338**, 207-215 (2004).
5. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28-36 (1994).
6. Mahony, S., Auron, P.E. & Benos, P.V. DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput Biol* **3**, e61 (2007).
7. Cha, J.Y., Kim, H., Kim, K.S., Hur, M.W. & Ahn, Y. Identification of transacting factors responsible for the tissue-specific expression of human glucose transporter type 2 isoform gene. Cooperative role of hepatocyte nuclear factors 1alpha and 3beta. *J Biol Chem* **275**, 18358-18365 (2000).
8. Gauthier, B.R. *et al.* Hepatic nuclear factor-3 (HNF-3 or FOXA2) regulates glucagon gene transcription by binding to the G1 and G2 promoter elements. *Mol Endocrinol* **16**, 170-183 (2002).
9. Hashimoto, T. *et al.* Regulation of ATP-sensitive potassium channel subunit Kir6.2 expression in rat intestinal insulin-producing progenitor cells. *J Biol Chem* **280**, 1893-1900 (2005).
10. Lantz, K.A. *et al.* FOXA2 regulates multiple pathways of insulin secretion. *J Clin Invest* **114**, 512-520 (2004).
11. Raum, J.C. *et al.* FOXA2, Nkx2.2, and PDX-1 regulate islet beta-cell-specific mafA expression through conserved sequences located between base pairs -8118 and -7750 upstream from the transcription start site. *Mol Cell Biol* **26**, 5735-5743 (2006).
12. Samaras, S.E. *et al.* Conserved sequences in a tissue-specific regulatory region of the pdx-1 gene mediate transcription in Pancreatic beta cells: role for hepatocyte nuclear factor 3 beta and Pax6. *Mol Cell Biol* **22**, 4702-4713 (2002).
13. Watada, H., Scheel, D.W., Leung, J. & German, M.S. Distinct gene expression programs function in progenitor and mature islet cells. *J Biol Chem* **278**, 17130-17140 (2003).
14. Keller, D.M. *et al.* Characterization of pancreatic transcription factor Pdx-1 binding sites using promoter microarray and serial analysis of chromatin occupancy. *J Biol Chem* **282**, 32084-32092 (2007).
15. Carty, M.D., Lillquist, J.S., Peshavaria, M., Stein, R. & Soeller, W.C. Identification of cis- and trans-active factors regulating human islet amyloid

polypeptide gene expression in pancreatic beta-cells. *J Biol Chem* **272**, 11986-11993 (1997).
16. German, M. *et al.* The insulin gene promoter. A simplified nomenclature. *Diabetes* **44**, 1002-1004 (1995).
17. Gerrish, K., Cissell, M.A. & Stein, R. The role of hepatic nuclear factor 1 alpha and PDX-1 in transcriptional regulation of the pdx-1 gene. *J Biol Chem* **276**, 47775-47784 (2001).
18. Leonard, J. *et al.* Characterization of somatostatin transactivating factor-1, a novel homeobox factor that stimulates somatostatin expression in pancreatic islet cells. *Mol Endocrinol* **7**, 1275-1283 (1993).
19. Shelton, K.D., Franklin, A.J., Khoor, A., Beechem, J. & Magnuson, M.A. Multiple elements in the upstream glucokinase promoter contribute to transcription in insulinoma cells. *Mol Cell Biol* **12**, 4578-4589 (1992).
20. Waeber, G., Thompson, N., Nicod, P. & Bonny, C. Transcriptional activation of the GLUT2 gene by the IPF-1/STF-1/IDX-1 homeobox factor. *Mol Endocrinol* **10**, 1327-1334 (1996).
21. Reich, M., et al. (2006) GenePattern 2.0. *Nature Genetics*, **38**, 500–50122.
22. Odom, D.T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39**, 730-732 (2007).
23. Robertson, A.G. et al. Genome wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res* (2008).