Integrated Epigenetic Mapping of Human and Mouse Salivary Gene Regulation

D.G. Michael, T.J.F. Pranzatelli, B.M. Warner, H. Yin, and J.A. Chiorini

Appendix

Materials and Methods:

Animals: All animal procedures were performed using female 8-week-old BALB/c mice (*n* = 2) acquired from Jackson Labs. All mouse studies were conducted in an AAALAC-accredited facility following the Institutional Animal Care and Use Committee Protocol (NIDCR).

RNA expression analyses: Identification of highly expressed and enriched transcription factors was performed via analysis of normalized RNA-seq data from thirty-seven human tissues provided by the Human Protein Atlas Version 18 (ProteinAtlas.org/about/download) and the BioGPS Mouse GeneAtlas V3 from ninety-six mouse tissues (GEO Accession GSE10246) (Uhlen et al. 2015; Wu et al. 2016). For each dataset, data was parsed and analyzed using Python3 scripts using Pandas, SKLearn, SciPy, Numpy, Matplotlib, Jupyter Notebook and Seaborn computational libraries (Kluyver et al. 2016; Krzywinski et al. 2009; van der Walt et al. 2011). Within each dataset, a distribution of each gene's expression across all assayed tissues was constructed and used to calculate z-scores for expression enrichment. Identification and analysis of transcription factor expression was performed using annotations acquired from the Animal TFDB, a database which annotates proteins with known or predicted DNA binding domains(Zhang et al. 2015). Spearman rank-based correlation visualization and clustering of transcription factor expression patterns across all tissues was performed on the top 50% of all expressed transcription factors to minimize the impact of rank fluctuations within lowly expressed genes. Comparison of mouse and human orthologue expression ranks was performed using Python3 and Ensembl orthologue mappings for hg19 and mm9 genome builds (ensembl.org/).

DNase1 Digital Genomic Footprinting: Identification of chromatin state via DNase1 Digital Genomic Footprinting was performed according to protocols previously described (Hesselberth et al. 2009; Sabo et al. 2006). Submandibular salivary glands were extracted from two BALB/c female mice and nuclei isolated using a GentleMACS dissociator (Miltenyi Biotec, Auburn CA) in a sucrose buffer (250 mM D-Sucrose, 10 mM Tris-HCl pH 7.5, 1 mM MgCl2) prior to filtering through a 100 uM filter. DNase1 treatment was performed using the two-hit method of Hesselberth et al in which the suspended nuclei were mixed with a 2x reaction DNase1 reaction buffer containing titrated DNase1 at concentrations of 20, 15, 10, 7.5, 5, 3.75, 2.5, 1.875 and 0 U / ml. Samples were incubated for 3 minutes prior to addition of a stop solution (50 mM Tris-HCl (pH 8.0), 100 mM NaCl, 0.1% SDS, 100 mM EDTA (pH 8.0), 10 μg/ml Ribonuclease A, 1 mM spermidine, 0.3 mM spermine) prior to addition of 20 ug/mL Proteinase K and overnight incubation at 55°C. DNase1 hypersensitivity libraries were constructed as described by

Hesselberth et al (Hesselberth et al. 2009) and sequenced on an Illumina HiSeq 2500 to a read depth of 163,447,713 and 169,978,640 reads per biological replicate.

Controlling for DNase1 hypersensitivity biases:
DNase1 has been noted to exhibit sequence specific biases that convolute transcription factor footprints with natural enzymatic specificities. To control for these biases, we utilized the HINT-BC DNase1 footprinting package which calculates DNase1 bias based on data generated from deproteinized DNA and uses this information to remove the instrinsic DNase1 cleavage signal from the DNase1 data. HINT-BC was called using the pre-computed '-default-bias-correction' argument for analyses used within this paper(Gusmao et al. 2016).

Bivariate Genomic Footprinting Analysis: DNase1 digital genomic footprinting data was downloaded from ENCODE data repository (Dunham et al. 2012). To detect changes in transcription factor activity between salivary gland, lung, and heart tissues we utilized the BaGFoot algorithm to perform bivariate genomic footprinting analysis(Baek et al. 2017). BaGfoot data produces a quantitative readout on the altered chromatin accessibility and footprint depth for each PWM (position weight matrice). To enable rank-ordering, we summarized these two dimensions of information using an equation that selects PWMs which have high levels of chromatin accessibility and increases in footprint depth (equation 1). In this equation, theta represents the deviation of each point from the optimal 45° angle, FPD represents the BaGfoot differential footprint depth and CA represents differential chromatin accessibility between the two states (equation 1):

(1) $TFscore = \sqrt{FPD^2 + CA^2} \times (\frac{180 - \theta}{180})$

The bivariate genomic footprinting algorithm (BaGfoot) described by the Hager lab at NIH/NCI publication *"Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity."* provides additional details on the analytical process we used (Baek et al, 2017). Using this method, we reconstructed TF activity levels via comparison of motif features across three whole genome DNase1-seq datasets (SG v. lung, SG v. heart).

Position Weight Matrices (PWMs):
PWMs are a mathematical representation of transcription factor binding motifs. A PWM is generated from analyses which aggregate the genome-wide or *in vitro* binding specificity of a transcription factor across many different sequences. Each position in a PWM represents the preferred nucleotide at a given position in a motif, log-normalized for the frequency of that nucleotide within the relevant genome. For additional literature on PWMs, we recommend the text "Introduction to Protein-DNA interactions" by Dr. Gary Stormo.

Genome Wide Gene Regulatory Network Reconstruction: Following sequencing and generation of FASTQ files, reads were aligned to the MM9 mouse genome using Bowtie2(Langmead and Salzberg 2012). Regions of enriched read density were determined using the HOMER FindPeaks algorithm and TF footprints were detected using HINT-BC after correction for known DNase1 cleavage biases(Gusmao et al. 2016). For each detected footprint, a prediction of TF

occupancy was generated by scanning TF PWMs across the MM9 genome using FIMO at a p-value cut-off of *p* = 1E-04. PWM scanning was performed using 722 PWMs for mouse TFs downloaded from the CIS-BP PWM database and the MEME suite(Grant et al. 2011; Weirauch et al. 2014). In situations where more than one PWM overlapped with a footprint, the PWM that minimized the p-value for each footprint was selected as the best hit for visualization and analysis.

FANTOM5 CAGE eRNA (enhancer RNA) expression data for the mouse was downloaded and used to annotate distal regulatory elements(Andersson et al. 2014). Association of distal enhancer regions with predicted promoter targets was performed by identifying statistically significant correlations between all gene-enhancer pairs across one million base pair windows with a false-discovery threshold of 0.05. False discovery rates (FDR) for enhancer-promoter pairing were estimated from an empirical null distribution generated by sampling with replacement from the approximately 1,200,000,000 possible enhancer-promoter combinations.

There are several methods for selection of promoter intervals. We used the well-accepted approach for promoter definition published by Neph et al. who has developed human transcriptional regulatory networks from DNase1-seq. Under this approach, a five kilobase window is defined as putatively regulatory and used for downstream analyses (Neph et al. 2012). In the current manuscript, we improved on this approach by also including well-defined, FANTOM 5 CAGE-seq supported distal regulatory regions as an additional regulatory interval.

To reconstruct the mouse salivary gland gene regulatory network, all TF footprints falling within a five kilobase window of a known transcription start site for the mm9 mouse genome reference build were annotated as predicted regulators of that gene. Footprints falling within detected enhancer regions were also annotated as predicted regulators for any significantly correlated enhancer. These TF-to-gene relationships were used to construct a gene regulatory network summarizing detected regulatory relationships within the DNase1 footprinting data.

Identification of enriched regulatory relationships within the mouse salivary gland gene regulatory network was performed using Python3 and the ENRICHR Gene Ontology Enrichment analysis API(Kuleshov et al. 2016). For each TF with a known PWM, all gene targets detected within the reconstructed gene regulatory network were queried against the KEGG 2016 database to identify significantly enriched TF predicted to bind the promoters of genes within known pathways (FDR < 0.05). Statistically significant interactions were visualized using the CIRCOS Perl library(Krzywinski et al. 2009). Although the TF network contained all known transcription factors, the top 20 transcription factors weighted by TF expression rank (top 1-20), salivary tissue specificity (z-score maximized ranks 1-20) and normalized bivariate genomic footprinting activity (ranks 1-20) were selected for CIRCOS visualization and additional analysis. A detailed examination of the code and the impact of computational parameters on network inference accuracy is available in the ATAC2GRN manuscript recently published by Pranzatelli et al. (Pranzatelli et al. 2018).

Histology and Immunohistochemistry: Salivary gland protein expression confirmation of RNA expression and DNase-1 activity-prioritized TF was assessed using publicly-available immunohistochemical digital whole slide images from Human Protein Atlas (HPA). When prioritized TF immunohistochemistry was not available through HPA, immunofluorescence using pre-validated polyclonal rabbit antibodies (XBP1, NR4A1, ELF5, ETV1, PLAG1; Atlas Antibodies, Sigma-Aldrich, St. Louis Missouri) was performed using standard techniques and the manufacturer's recommended antibody dilution concentrations on normal human submandibular gland tissues procured from the Human Cooperative Tissue Network. For each, a semi-quantitative score (0, absent; 1, weak; 2, moderate; 3, strong) for acini and ducts with subcellular localization to nuclei and/or cytoplasm was performed by a board certified oral and maxillofacial pathologist (BMW). Salivary gland immunohistochemistry for (HPA TF) was downloaded from the Human Protein Atlas Version 18 (ProteinAtlas.org)(Uhlen et al. 2015). Analysis and visualization of immunohistochemistry quantification data was performed in Python3.

Additional citations for supplemental methods:

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T et al. 2014. An atlas of active enhancers across human cell types and tissues. Nature. 507(7493):455-+.

Baek S, Goldstein I, Hager GL. 2017. Bivariate genomic footprinting detects changes in transcription factor activity. Cell Rep. 19(8):1710-1722.

Grant CE, Bailey TL, Noble WS. 2011. Fimo: Scanning for occurrences of a given motif. Bioinformatics. 27(7):1017-1018.

Gusmao EG, Allhoff M, Zenke M, Costa IG. 2016. Analysis of computational footprinting methods for dnase sequencing experiments. Nat Methods. 13(4):303-309.

Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS et al. 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods. 6(4):283-289.

Kluyver T, Ragan-Kelley B, Perez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S et al. 2016. Jupyter notebooks-a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas.87-90.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. Genome Res. 19(9):1639-1645.

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A et al. 2016. Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 44(W1):W90-97.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with bowtie 2. Nat Methods. 9(4):357-U354.

Pranzatelli TJF, Michael DG, Chiorini JA. 2018. Atac2grn: Optimized atac-seq and dnase1-seq pipelines for rapid and accurate genome regulatory network inference. Bmc Genomics. 19(1):563.

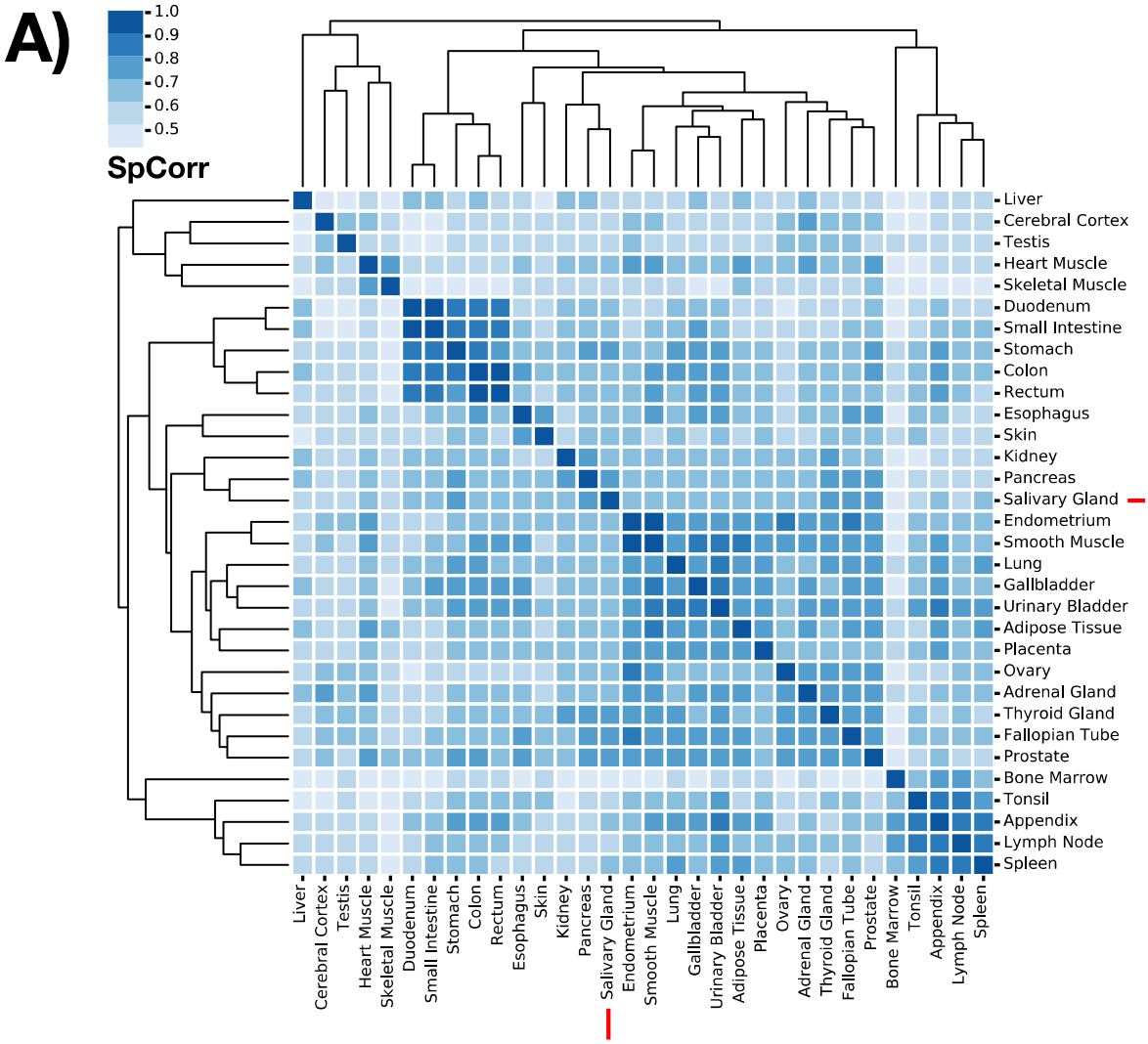
Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Hua C, Man Y, Rosenzweig E, Goldy J, Haydock A et al. 2006. Genome-scale mapping of dnase i sensitivity in vivo using tiling DNA microarrays. Nat Methods. 3(7):511-518.

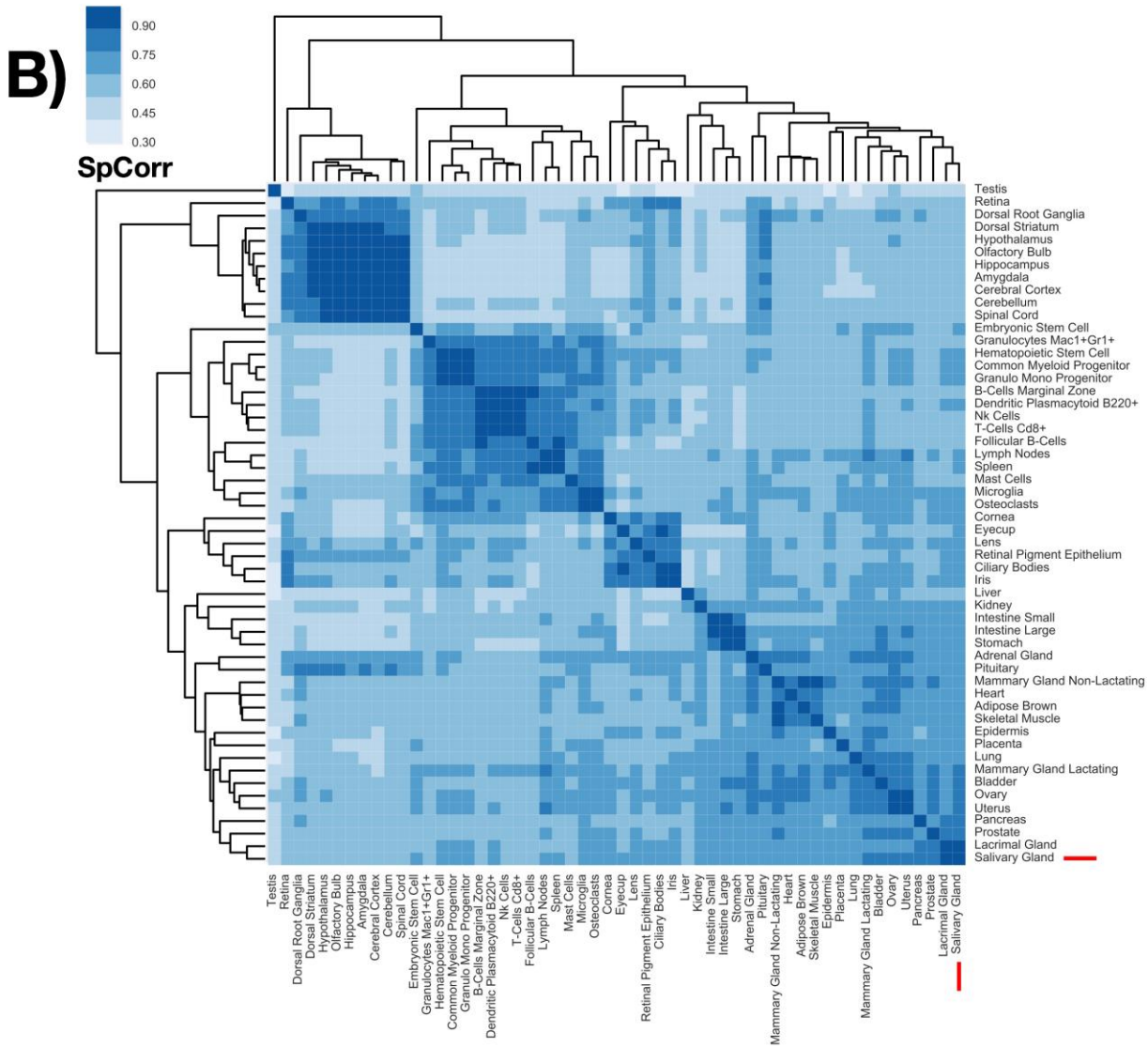
Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A et al. 2015. Tissue-based map of the human proteome. Science. 347(6220).

van der Walt S, Colbert SC, Varoquaux G. 2011. The numpy array: A structure for efficient numerical computation. Comput Sci Eng. 13(2):22-30.

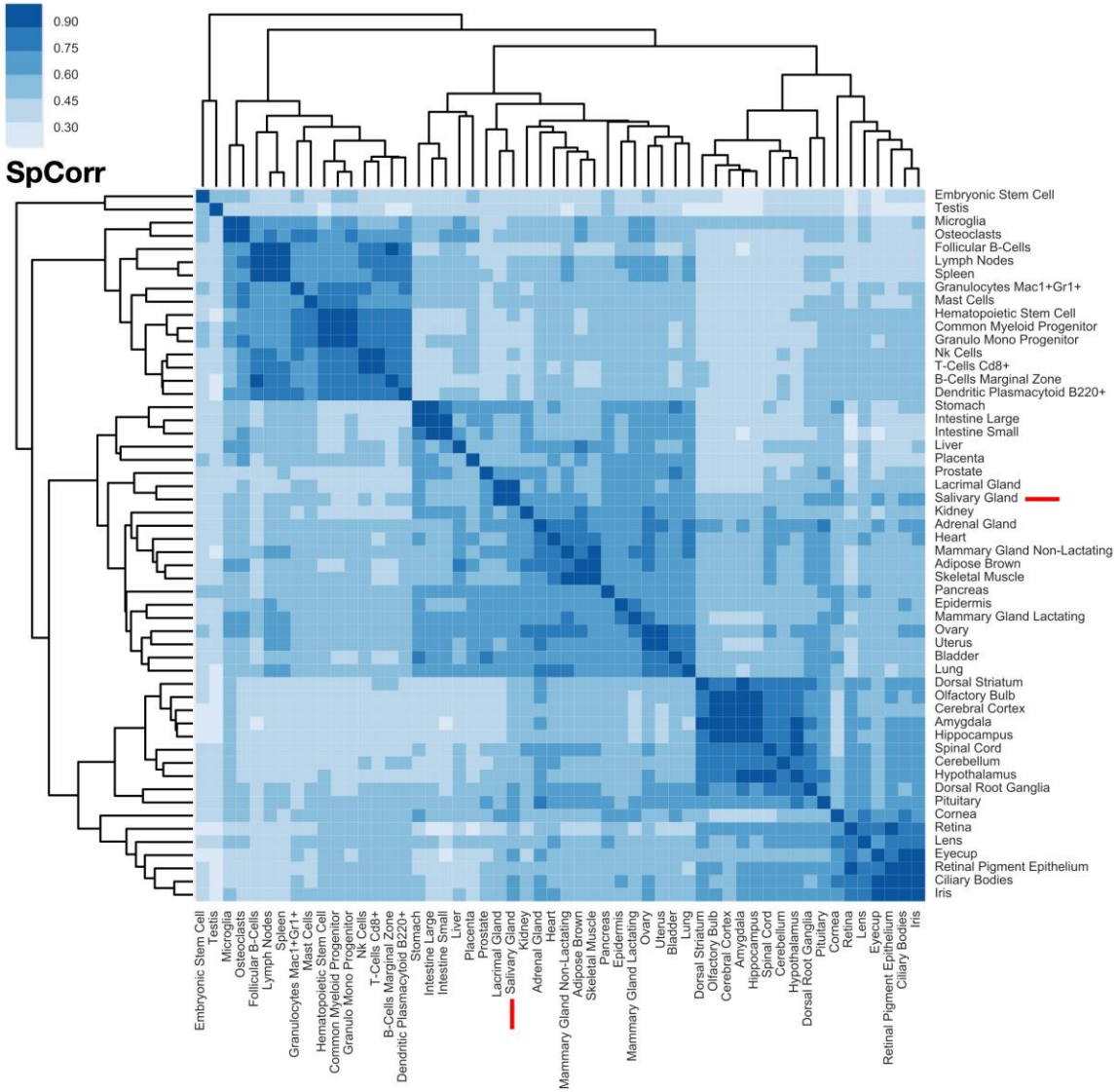
Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K et al. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 158(6):1431-1443.

Zhang HM, Liu T, Liu CJ, Song SY, Zhang XT, Liu W, Jia HB, Xue Y, Guo AY. 2015. Animaltfdb 2.0: A resource for expression, prediction and functional study of animal transcription factors. Nucleic Acids Res. 43(D1):D76-D81.


A)
1.0
0.9
0.8
0.7
0.6
0.5
SpCorr
Liver
Cerebral Cortex
Testis
Heart Muscle
Skeletal Muscle
Duodenum
Small Intestine
Stomach
Colon
Rectum
Esophagus
Skin
Kidney
Pancreas
Salivary Gland
Endometrium
Smooth Muscle
Lung
Gallbladder
Urinary Bladder
Adipose Tissue
Placenta
Ovary
Adrenal Gland
Thyroid Gland
Fallopian Tube
Prostate
Bone Marrow
Tonsil
Appendix
Lymph Node
Spleen


B)
0.90
0.75
0.60
0.45
0.30
SpCorr
Testis
Retina
Dorsal Root Ganglia
Dorsal Striatum
Hypothalamus
Olfactory Bulb
Hippocampus
Amygdala
Cerebral Cortex
Cerebellum
Spinal Cord
Embryonic Stem Cell
Granulocytes Mac1+Gr1+
Hematopoietic Stem Cell
Common Myeloid Progenitor
Granulo Mono Progenitor
B-Cells Marginal Zone
Dendritic Plasmacytoid B220+
Nk Cells
T-Cells Cd8+
Follicular B-Cells
Lymph Nodes
Spleen
Mast Cells
Microglia
Osteoclasts
Cornea
Eyecup
Lens
Retinal Pigment Epithelium
Ciliary Bodies
Iris
Liver
Kidney
Intestine Small
Intestine Large
Stomach
Adrenal Gland
Pituitary
Mammary Gland Non-Lactating
Heart
Adipose Brown
Skeletal Muscle
Epidermis
Placenta
Lung
Mammary Gland Lactating
Bladder
Ovary
Uterus
Pancreas
Prostate
Lacrimal Gland
Salivary Gland

C)



Appendix Figure: Spearman rank correlations across the mouse and human gene atlases. (A) Spearman rank correlations of all gene types, across all tissues within the Human Protein Atlas. (B) Spearman rank correlations of all gene types across all tissues within the BioGPS MOE430 mouse atlas. (C) Spearman rank correlations of the top 50% of all TFs across all tissues within the BioGPS MOE430 mouse atlas.

TF name	SG - Lung Activity	SG - Heart Activity	Lung - Heart Activity
Plag1	0.76516977	0.80124418	0
Egr1_3	0.72405601	0.594958	0
Zfp161	0.57021331	0.13200554	0
Zfp740	0.56912465	0.44183499	0
Tfap2a_2	0.55084986	0.27833959	0
E2f2	0.50270273	0.13640001	0
Rreb1	0.50198449	0.58498463	0
Nrf1_1	0.48310947	0.05848337	0
Tcfap2e	0.47962993	0.26311493	0
Plagl1	0.46184682	0.16604825	0
Sp4	0.46057447	0.34659287	0
E2f3	0.45849807	0.11443655	0
Sp1_2	0.45711077	0.28334977	0
Zic3	0.42938955	0.10852828	0
Zic2	0.42651726	0.10797199	0
Zic1	0.42416981	0.10237015	0
Tfap2a_1	0.40494543	0.22461792	0
Nrf1_2	0.40052754	0	0
Zfp281	0.40038014	0.22646609	0
Tcfap2c	0.39938116	0.24336959	0
Zbtb33	0.39871866	0.04905432	0
Gcm1	0.38612217	0.24046205	0
Tcfap2b	0.3825005	0.2135228	0
Ap2alpha	0.3677217	0.22976479	0
E2f7	0.36724116	0.12243952	0
E2f1_2	0.36656694	0.12566404	0
E2f1_1	0.36628996	0.08577499	0
Ap2gamma	0.35207572	0.22611865	0
Zfx_2	0.32791775	0.13621596	0
Egr1_1	0.315268	0.12386824	0
p53_1	0.30784288	0.16492986	0
Rest	0.28877987	0.16721731	0
Ascl2	0.28739159	0.11643973	0
Mtf1	0.28232306	0.1137039	0
Pax5_3	0.28169736	0.09739712	0

Sp100	0.27991597	0.37958406	0.08131261
Pax5_2	0.27657963	0.14617274	0
Egr1 4	0.27079998	0.10089031	0
E2f4	0.25717708	0.04964115	0
n-Myc_2	0.25681613	0.07710345	0
E2f6	0.25557548	0.07381888	0
E2A:PU.1	0.25206939	0.40965681	0.16467505
Rest:Nrsf	0.25004248	0.15306109	0
BORIS	0.24650508	0.04168141	0
Egr1_2	0.2440733	0.05031913	0
Tcfe2a	0.24075229	0.29052944	0.05672062
ESR1	0.24051599	0.13182275	0
Tcfcp2l1_2	0.23947421	0.1266892	0
Glis2	0.23427118	0.01907784	0
Sp1_1	0.23273508	0	0
Myc_3	0.23169854	0.08598799	0
NHLH1	0.22207596	0.04499839	0
Hif1a:Arnt	0.2161769	0.14242521	0
NFKB1	0.21593248	0.08701918	0
E2f1_3	0.2136279	0.02891661	0
Tcfcp2l1_1	0.21277168	0.15797434	0
Egr2	0.20653515	0	0
Zbtb7b	0.20233859	0	0
Zbtb3	0.20207352	0.11173127	0
Myf	0.20198965	0.09936002	0
Pax8	0.19537344	0.11052804	0
Klf4_2	0.19186089	0	0
Smad4	0.19143423	0.04420152	0
NFkB-p50,p52	0.18910963	0.05988975	0
MyoD	0.18906072	0	0
Prdm9	0.18165333	0.08351386	0
Insm1	0.17913028	0.05801215	0
Nf1_1	0.17438729	0	0
Hif2a	0.17332597	0.0608262	0
Pax5_1	0.16932244	0.0639682	0
Klf4_1	0.16743941	0	0
Hif-1a	0.16228256	0.04821955	0
HNF6	0.15973402	0.21790336	0.19130324
Pax2	0.15953878	0	0

Pax4_2	0.15601209	0.33351593	0.04655165
YY1_1	0.15474342	0	0
Hic1	0.15147129	0.17403046	0.12319264
Ebf1_2	0.15057623	0	0
Klf7	0.14117496	0	0
Zfp423	0.14019125	0.05350816	0
Gli3	0.13750378	0.07629299	0
CTCF_1	0.13637786	0	0
Myc_2	0.13414104	0.05523112	0
CTCF_2	0.13229037	0	0
p63	0.13163415	0.13740231	0.00801381
Ebf1_1	0.13126659	0	0
Gmeb1	0.13047775	0	0
Zfp128	0.12866008	0.08659145	0
Tcf12	0.12804183	0	0
TLX1::NFIC	0.12690329	0	0
E2A	0.12445917	0	0
Zfx_1	0.12217073	0.03570872	0
Ebf1_3	0.12160191	0	0
SCL	0.12090491	0	0
Myc_1	0.11612543	0	0
n-Myc_1	0.11483211	0	0
Mizf	0.11112069	0	0
Smad2	0.10902559	0	0
Fxr	0.10874425	0.08120784	0
Spdef_1	0.10571409	0	0
ESR2	0.10357711	0.05980391	0
Tlx?	0.10172765	0	0
EBNA1	0.1004287	0.03886248	0
NeuroD1	0.10034291	0	0
Atoh1	0.09943244	0	0
HEB?	0.09400495	0.01682886	0
Rarg	0.09344855	0.07493769	0.14536201
Max_1	0.09304514	0	0
Max_3	0.09204845	0	0
Tbx5	0.09195309	0.17689159	0.17341703
Myb_2	0.0902439	0	0
p53_3	0.08697644	0.09999548	0.01612857
Myf6	0.08404867	0.03875249	0

NFkB-cRel-p50	0.08369421	0	0
NFkB	0.08301263	0	0
Rxra_1	0.08295232	0	0
Ewsr1:Fli1	0.07843752	0	0
PPARG::RXRA	0.07746193	0	0
AR-halfsite	0.07284064	0.03910309	0
Mafb_2	0.07127694	0	0
Rxr:Rar:Dr5	0.07113531	0.0445766	0.09975778
Srebp2	0.06824014	0	0
VDR	0.06772075	0.05653844	0.06125318
Elk1_2	0.06588773	0	0
Hnf4a_3	0.06365899	0	0
Myf5	0.06309971	0	0
MyoG	0.06132212	0	0
Olig2	0.05990946	0	0
Smad3_2	0.05987962	0.02969978	0
Gsc_2327.3	0.056451	0.1688219	0.14601403
Myc:Max	0.05632188	0	0
Tbox:Smad	0.05419762	0.02576221	0
Obox3	0.05062985	0.07602875	0.11976692
Myb_1	0.04584717	0	0
Znf143	0.04092898	0	0
Six4	0.04066355	0	0.00511032
Bmyb	0.04048675	0	0
ZNF143|STAF	0.03815918	0	0
Pitx3	0.03787302	0	0
GATA:SCL	0.03689351	0	0
CTCF-SatelliteElement	0.03568081	0	0
Elk4_1	0.02573439	0	0
Obox5_1	0.01464844	0.08479778	0.11356475

Appendix Table: Normalized BagFoot activity levels for (salivary gland - heart) and (salivary gland - lung) comparisons for all TFs with > 0 normalized activity in the salivary glands.