Sequence and chromatin determinants of cell-type specific transcription factor binding: supplementary data

Aaron Arvey[1], Phaedra Agius[1], William Stafford Noble[2], and Christina Leslie[1*]
[1]Computational Biology Program, Memorial Sloan-Kettering Cancer Center, New York, NY
[2]Department of Genome Sciences, University of Washington, Seattle, WA

March 15, 2012

Table S1: List of all TF ChIP-seq experiments analyzed in the study.

File name	Cell	TF
wgEncodeYaleChIPseqRawDataRep1Helas3Ap2alpha	Helas3	AP2A1
wgEncodeYaleChIPseqRawDataRep2Helas3Ap2alpha	Helas3	AP2A1
wgEncodeYaleChIPseqRawDataRep1Helas3Ap2gamma	Helas3	TFAP2C
wgEncodeYaleChIPseqRawDataRep2Helas3Ap2gamma	Helas3	TFAP2C
wgEncodeYaleChIPseqRawDataRep1K562Atf3	K562	ATF3
wgEncodeYaleChIPseqRawDataRep2K562Atf3	K562	ATF3
wgEncodeYaleChIPseqRawDataRep1Helas3Baf155Musigg	Helas3	SMARCC1
wgEncodeYaleChIPseqRawDataRep2Helas3Baf155Musigg	Helas3	SMARCC1
wgEncodeYaleChIPseqRawDataRep1Helas3Baf170Musigg	Helas3	SMARCC2
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Batf	Gm12878	BATF
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Batf	Gm12878	BATF
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Bcl11a	Gm12878	BCL11A
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Bcl11a	Gm12878	BCL11A
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Bcl3Pcr1xBcl3	Gm12878	BCL3
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Bcl3Pcr1xBcl3	Gm12878	BCL3
wgEncodeYaleChIPseqRawDataRep1Helas3Bdp1	Helas3	BDP1
wgEncodeYaleChIPseqRawDataRep2Helas3Bdp1	Helas3	BDP1
wgEncodeYaleChIPseqRawDataRep1K562Bdp1	K562	BDP1
wgEncodeYaleChIPseqRawDataRep2K562Bdp1	K562	BDP1
wgEncodeYaleChIPseqRawDataRep3K562Bdp1	K562	BDP1
wgEncodeYaleChIPseqRawDataRep1Helas3Brf1	Helas3	BRF1
wgEncodeYaleChIPseqRawDataRep1K562Brf1	K562	BRF1
wgEncodeYaleChIPseqRawDataRep2K562Brf1	K562	BRF1
wgEncodeYaleChIPseqRawDataRep3K562Brf1	K562	BRF1
wgEncodeYaleChIPseqRawDataRep1K562Brf2	K562	BRF2
wgEncodeYaleChIPseqRawDataRep3K562Brf2	K562	BRF2
(continued...)		

*Corresponding author, cleslie@cbio.mskcc.org

Table S1: List of all TF ChIP-seq experiments analyzed in the study.

File name	Cell	TF
wgEncodeYaleChIPseqRawDataRep1Helas3Brg1Iggmus	Helas3	BRG1
wgEncodeYaleChIPseqRawDataRep2Helas3Brg1Iggmus	Helas3	BRG1
wgEncodeYaleChIPseqRawDataRep1K562Brg1Musigg	K562	BRG1
wgEncodeYaleChIPseqRawDataRep2K562Brg1Musigg	K562	BRG1
wgEncodeYaleChIPseqRawDataRep1Gm12878Cfos	Gm12878	FOS
wgEncodeYaleChIPseqRawDataRep2Gm12878Cfos	Gm12878	FOS
wgEncodeYaleChIPseqRawDataRep3Gm12878Cfos	Gm12878	FOS
wgEncodeYaleChIPseqRawDataRep1Helas3Cfos	Helas3	FOS
wgEncodeYaleChIPseqRawDataRep2Helas3Cfos	Helas3	FOS
wgEncodeYaleChIPseqRawDataRep1K562Cfos	K562	FOS
wgEncodeYaleChIPseqRawDataRep2K562Cfos	K562	FOS
wgEncodeYaleChIPseqRawDataRep3K562Cfos	K562	FOS
wgEncodeYaleChIPseqRawDataRep1Gm12878Cjun	Gm12878	JUN
wgEncodeYaleChIPseqRawDataRep2Gm12878Cjun	Gm12878	JUN
wgEncodeYaleChIPseqRawDataRep1Helas3CjunRabigg	Helas3	JUN
wgEncodeYaleChIPseqRawDataRep2Helas3CjunRabigg	Helas3	JUN
wgEncodeYaleChIPseqRawDataRep1K562Cjun	K562	JUN
wgEncodeYaleChIPseqRawDataRep2K562Cjun	K562	JUN
wgEncodeYaleChIPseqRawDataRep1Helas3Cmyc	Helas3	MYC
wgEncodeYaleChIPseqRawDataRep2Helas3Cmyc	Helas3	MYC
wgEncodeYaleChIPseqRawDataRep1K562Cmyc	K562	MYC
wgEncodeYaleChIPseqRawDataRep2K562Cmyc	K562	MYC
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Control	Gm12878	Input-Control
wgEncodeHudsonalphaChipSeqRawDataRep1K562Control	K562	Input-Control
wgEncodeUwChIPSeqRawDataRep1Gm12878Ctcf	Gm12878	CTCF
wgEncodeUwChIPSeqRawDataRep2Gm12878Ctcf	Gm12878	CTCF
wgEncodeYaleChIPseqRawDataRep1Helas3E2f1	Helas3	E2F1
wgEncodeYaleChIPseqRawDataRep2Helas3E2f1	Helas3	E2F1
wgEncodeYaleChIPseqRawDataRep1Helas3E2f4	Helas3	E2F4
wgEncodeYaleChIPseqRawDataRep2Helas3E2f4	Helas3	E2F4
wgEncodeYaleChIPseqRawDataRep1K562bE2f4	K562	E2F4
wgEncodeYaleChIPseqRawDataRep2K562bE2f4	K562	E2F4
wgEncodeYaleChIPseqRawDataRep1Helas3E2f6	Helas3	E2F6
wgEncodeYaleChIPseqRawDataRep2Helas3E2f6	Helas3	E2F6
wgEncodeYaleChIPseqRawDataRep1K562bE2f6	K562	E2F6
wgEncodeYaleChIPseqRawDataRep2K562bE2f6	K562	E2F6
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Egr1Pcr2x	Gm12878	EGR1
wgEncodeHudsonalphaChipSeqRawDataRep1K562Egr1Pcr2x	K562	EGR1
wgEncodeHudsonalphaChipSeqRawDataRep2K562Egr1Pcr2x	K562	EGR1
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878GabpPcr2x	Gm12878	GABPA
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878GabpPcr2x	Gm12878	GABPA
wgEncodeHudsonalphaChipSeqRawDataRep1Helas3GabpPcr1x	Helas3	GABPA
(continued...)		

Table S1: List of all TF ChIP-seq experiments analyzed in the study.

File name	Cell	TF
wgEncodeHudsonalphaChipSeqRawDataRep2Helas3GabpPcr1x	Helas3	GABPA
wgEncodeHudsonalphaChipSeqRawDataRep1K562GabpPcr2x	K562	GABPA
wgEncodeHudsonalphaChipSeqRawDataRep2K562GabpPcr2x	K562	GABPA
wgEncodeYaleChIPseqRawDataRep1K562bGata1	K562	GATA1
wgEncodeYaleChIPseqRawDataRep2K562bGata1	K562	GATA1
wgEncodeYaleChIPseqRawDataRep1K562bGata2	K562	GATA2
wgEncodeYaleChIPseqRawDataRep2K562bGata2	K562	GATA2
wgEncodeYaleChIPseqRawDataRep1K562Gtf2b	K562	GTF2B
wgEncodeYaleChIPseqRawDataRep2K562Gtf2b	K562	GTF2B
wgEncodeYaleChIPseqRawDataRep1Helas3Ha	Helas3	HA-E2F1
wgEncodeYaleChIPseqRawDataRep2Helas3Ha	Helas3	HA-E2F1
wgEncodeHudsonalphaChipSeqRawDataRep1K562Hey1	K562	HEY1
wgEncodeHudsonalphaChipSeqRawDataRep2K562Hey1	K562	HEY1
wgEncodeYaleChIPseqRawDataRep1K562MusiggMusigg	K562	mouse-IGG
wgEncodeYaleChIPseqRawDataRep2K562MusiggMusigg	K562	mouse-IGG
wgEncodeYaleChIPseqRawDataRep1Helas3Ini1Musigg	Helas3	SMARCB1
wgEncodeYaleChIPseqRawDataRep2Helas3Ini1Musigg	Helas3	SMARCB1
wgEncodeYaleChIPseqRawDataRep1K562Ini1Musigg	K562	SMARCB1
wgEncodeYaleChIPseqRawDataRep2K562Ini1Musigg	K562	SMARCB1
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Irf4	Gm12878	IRF4
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Irf4	Gm12878	IRF4
wgEncodeYaleChIPseqRawDataRep1Gm12878Jund	Gm12878	JUND
wgEncodeYaleChIPseqRawDataRep2Gm12878Jund	Gm12878	JUND
wgEncodeYaleChIPseqRawDataRep1Helas3JundRabigg	Helas3	JUND
wgEncodeYaleChIPseqRawDataRep2Helas3JundRabigg	Helas3	JUND
wgEncodeYaleChIPseqRawDataRep1K562Jund	K562	JUND
wgEncodeYaleChIPseqRawDataRep2K562Jund	K562	JUND
wgEncodeYaleChIPseqRawDataRep2Helas3Largefragment	Helas3	Large-fragment
wgEncodeYaleChIPseqRawDataRep1Gm12878Max	Gm12878	MAX
wgEncodeYaleChIPseqRawDataRep2Gm12878Max	Gm12878	MAX
wgEncodeYaleChIPseqRawDataRep1Helas3Max	Helas3	MAX
wgEncodeYaleChIPseqRawDataRep2Helas3Max	Helas3	MAX
wgEncodeYaleChIPseqRawDataRep1K562Max	K562	MAX
wgEncodeYaleChIPseqRawDataRep2K562Max	K562	MAX
wgEncodeYaleChIPseqRawDataRep1Helas3Mouseigg	Helas3	mouse-IGG
wgEncodeYaleChIPseqRawDataRep2Helas3Mouseigg	Helas3	mouse-IGG
wgEncodeYaleChIPseqRawDataRep1K562Nelfe	K562	RDBP
wgEncodeYaleChIPseqRawDataRep2K562Nelfe	K562	RDBP
wgEncodeYaleChIPseqRawDataRep1K562Nfe2	K562	NFE2
wgEncodeYaleChIPseqRawDataRep2K562Nfe2	K562	NFE2
wgEncodeSydhTfbsGm12878Nfe2hStdRawDataRep1	Gm12878	NFE2
wgEncodeSydhTfbsGm12878Nfe2hStdRawDataRep2	Gm12878	NFE2
(continued...)		

Table S1: List of all TF ChIP-seq experiments analyzed in the study.

File name	Cell	TF
wgEncodeYaleChIPseqRawDataRep1Gm12878NfkbIggrab	Gm12878	NFKB
wgEncodeYaleChIPseqRawDataRep1Gm12878NfkbTnfa	Gm12878	NFKB
wgEncodeYaleChIPseqRawDataRep2Gm12878NfkbIggrab	Gm12878	NFKB
wgEncodeYaleChIPseqRawDataRep2Gm12878NfkbTnfa	Gm12878	NFKB
wgEncodeYaleChIPseqRawDataRep3Gm12878NfkbIggrab	Gm12878	NFKB
wgEncodeYaleChIPseqRawDataRep3Gm12878NfkbTnfa	Gm12878	NFKB
wgEncodeYaleChIPseqRawDataRep4Gm12878NfkbIggrab	Gm12878	NFKB
wgEncodeYaleChIPseqRawDataRep4Gm12878NfkbTnfa	Gm12878	NFKB
wgEncodeYaleChIPseqRawDataRep1K562Nfya	K562	NYFA
wgEncodeYaleChIPseqRawDataRep2K562Nfya	K562	NYFA
wgEncodeYaleChIPseqRawDataRep1K562Nfyb	K562	NYFB
wgEncodeYaleChIPseqRawDataRep2K562Nfyb	K562	NYFB
wgEncodeYaleChIPseqRawDataRep1Helas3Nrf1Musigg	Helas3	NRF1
wgEncodeYaleChIPseqRawDataRep2Helas3Nrf1Musigg	Helas3	NRF1
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878NrsfPcr2x	Gm12878	REST
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878NrsfPcr2x	Gm12878	REST
wgEncodeHudsonalphaChipSeqRawDataRep1K562NrsfPcr2x	K562	REST
wgEncodeHudsonalphaChipSeqRawDataRep2K562NrsfPcr2x	K562	REST
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878P300	Gm12878	EP300
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878P300	Gm12878	EP300
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Pax5c20	Gm12878	PAX5
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Pax5c20	Gm12878	PAX5
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Pbx3Pcr1x	Gm12878	PBX3
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Pbx3Pcr1x	Gm12878	PBX3
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Pol2Pcr2x	Gm12878	POLR2
wgEncodeYaleChIPseqRawDataRep1Gm12878Pol2	Gm12878	POLR2
wgEncodeYaleChIPseqRawDataRep1Gm12878Pol2Musigg	Gm12878	POLR2
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Pol2	Gm12878	POLR2
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Pol2Pcr2x	Gm12878	POLR2
wgEncodeYaleChIPseqRawDataRep2Gm12878Pol2	Gm12878	POLR2
wgEncodeYaleChIPseqRawDataRep2Gm12878Pol2Musigg	Gm12878	POLR2
wgEncodeYaleChIPseqRawDataRep3Gm12878Pol2Musigg	Gm12878	POLR2
wgEncodeYaleChIPseqRawDataRep4Gm12878Pol2Musigg	Gm12878	POLR2
wgEncodeYaleChIPseqRawDataRep5Gm12878Pol2Musigg	Gm12878	POLR2
wgEncodeYaleChIPseqRawDataRep6Gm12878Pol2Musigg	Gm12878	POLR2
wgEncodeYaleChIPseqRawDataRep7Gm12878Pol2Musigg	Gm12878	POLR2
wgEncodeHudsonalphaChipSeqRawDataRep1Helas3Pol2Pcr1x	Helas3	POLR2
wgEncodeYaleChIPseqRawDataRep1Helas3Pol2	Helas3	POLR2
wgEncodeYaleChIPseqRawDataRep2Helas3Pol2	Helas3	POLR2
wgEncodeHudsonalphaChipSeqRawDataRep1K562Pol2	K562	POLR2
wgEncodeYaleChIPseqRawDataRep1K562Pol2	K562	POLR2
wgEncodeYaleChIPseqRawDataRep1K562Pol2Musigg	K562	POLR2
(continued...)		

Table S1: List of all TF ChIP-seq experiments analyzed in the study.

File name	Cell	TF
wgEncodeHudsonalphaChipSeqRawDataRep2K562Pol2	K562	POLR2
wgEncodeYaleChIPseqRawDataRep2K562Pol2	K562	POLR2
wgEncodeYaleChIPseqRawDataRep2K562Pol2Musigg	K562	POLR2
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Pol24h8Pcr1x	Gm12878	POLR2
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Pol24h8Pcr1x	Gm12878	POLR2
wgEncodeHudsonalphaChipSeqRawDataRep1K562Pol24h8	K562	POLR2
wgEncodeHudsonalphaChipSeqRawDataRep2K562Pol24h8	K562	POLR2
wgEncodeYaleChIPseqRawDataRep1Gm12878Pol3	Gm12878	POLR3
wgEncodeYaleChIPseqRawDataRep2Gm12878Pol3	Gm12878	POLR3
wgEncodeYaleChIPseqRawDataRep1K562Pol3	K562	POLR3
wgEncodeYaleChIPseqRawDataRep2K562Pol3	K562	POLR3
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Pou2f2Pcr1x	Gm12878	POU2F2
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Pou2f2Pcr1x	Gm12878	POU2F2
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Pu1Pcr1x	Gm12878	SPI1
wgEncodeHudsonalphaChipSeqRawDataRep2K562Pu1	K562	SPI1
wgEncodeYaleChIPseqRawDataRep1Gm12878Rad21Iggrab	Gm12878	RAD21
wgEncodeYaleChIPseqRawDataRep2Gm12878Rad21Iggrab	Gm12878	RAD21
wgEncodeYaleChIPseqRawDataRep1K562Rad21	K562	RAD21
wgEncodeYaleChIPseqRawDataRep2K562Rad21	K562	RAD21
wgEncodeYaleChIPseqRawDataRep1Helas3Rpc155	Helas3	POLR3A
wgEncodeYaleChIPseqRawDataRep2Helas3Rpc155	Helas3	POLR3A
wgEncodeYaleChIPseqRawDataRep1K562Rpc155	K562	POLR3A
wgEncodeYaleChIPseqRawDataRep2K562Rpc155	K562	POLR3A
wgEncodeYaleChIPseqRawDataRep1K562bSetdb1	K562	SETDB1
wgEncodeYaleChIPseqRawDataRep1K562bSetdb1Mnase	K562	SETDB1
wgEncodeYaleChIPseqRawDataRep2K562bSetdb1	K562	SETDB1
wgEncodeYaleChIPseqRawDataRep2K562bSetdb1Mnase	K562	SETDB1
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Sin3ak20Pcr2x	Gm12878	SIN3A
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Sin3ak20Pcr2x	Gm12878	SIN3A
wgEncodeHudsonalphaChipSeqRawDataRep1K562Sin3ak20Pcr2x	K562	SIN3A
wgEncodeHudsonalphaChipSeqRawDataRep2K562Sin3ak20Pcr2x	K562	SIN3A
wgEncodeYaleChIPseqRawDataRep1K562Sirt6	K562	SIRT6
wgEncodeYaleChIPseqRawDataRep2K562Sirt6	K562	SIRT6
wgEncodeHudsonalphaChipSeqRawDataRep1K562Six5	K562	SIX5
wgEncodeHudsonalphaChipSeqRawDataRep2K562Six5	K562	SIX5
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Sp1Pcr1x	Gm12878	SP1
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Sp1Pcr1x	Gm12878	SP1
wgEncodeHudsonalphaChipSeqRawDataRep1K562Sp1Pcr1x	K562	SP1
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878SrfPcr2x	Gm12878	SRF
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878SrfPcr2x	Gm12878	SRF
wgEncodeHudsonalphaChipSeqRawDataRep1K562SrfPcr2x	K562	SRF
wgEncodeHudsonalphaChipSeqRawDataRep2K562SrfPcr2x	K562	SRF
(continued...)		

Table S1: List of all TF ChIP-seq experiments analyzed in the study.

File name	Cell	TF
wgEncodeYaleChIPseqRawDataRep1Helas3ifngStat1	Helas3ifng	STAT1
wgEncodeYaleChIPseqRawDataRep2Helas3ifngStat1	Helas3ifng	STAT1
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Taf1Pcr2x	Gm12878	TAF1
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Taf1Pcr1x	Gm12878	TAF1
wgEncodeHudsonalphaChipSeqRawDataRep1Helas3Taf1Pcr1x	Helas3	TAF1
wgEncodeHudsonalphaChipSeqRawDataRep2Helas3Taf1Pcr1x	Helas3	TAF1
wgEncodeHudsonalphaChipSeqRawDataRep1K562Taf1Pcr1x	K562	TAF1
wgEncodeHudsonalphaChipSeqRawDataRep2K562Taf1Pcr1x	K562	TAF1
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Tafii	Gm12878	TAF7
wgEncodeSydhTfbsGm12878TbpIggmusRawDataRep1	Gm12878	TBP
wgEncodeSydhTfbsGm12878TbpIggmusRawDataRep2	Gm12878	TBP
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Tcf12Pcr1x	Gm12878	TCF12
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Tcf12	Gm12878	TCF12
wgEncodeYaleChIPseqRawDataRep1Helas3Tfiiic	Helas3	GTF3C2
wgEncodeYaleChIPseqRawDataRep2Helas3Tfiiic	Helas3	GTF3C2
wgEncodeYaleChIPseqRawDataRep1K562Tfiiic	K562	GTF3C2
wgEncodeYaleChIPseqRawDataRep2K562Tfiiic	K562	GTF3C2
wgEncodeYaleChIPseqRawDataRep3K562Tfiiic	K562	GTF3C2
wgEncodeYaleChIPseqRawDataRep1Gm12878Tr4	Gm12878	NR2C2
wgEncodeYaleChIPseqRawDataRep2Gm12878Tr4	Gm12878	NR2C2
wgEncodeYaleChIPseqRawDataRep1Helas3Tr4	Helas3	NR2C2
wgEncodeYaleChIPseqRawDataRep2Helas3Tr4	Helas3	NR2C2
wgEncodeYaleChIPseqRawDataRep1K562bTr4	K562	NR2C2
wgEncodeYaleChIPseqRawDataRep2K562bTr4	K562	NR2C2
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Usf1Pcr2x	Gm12878	USF1
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Usf1Pcr2x	Gm12878	USF1
wgEncodeHudsonalphaChipSeqRawDataRep1K562Usf1Pcr2x	K562	USF1
wgEncodeHudsonalphaChipSeqRawDataRep2K562Usf1Pcr2x	K562	USF1
wgEncodeSydhTfbsGm12878WhipIggmusRawDataRep2	Gm12878	WRNIP1
wgEncodeYaleChIPseqRawDataRep1K562Xrcc4	K562	XRCC4
wgEncodeYaleChIPseqRawDataRep2K562Xrcc4	K562	XRCC4
wgEncodeYaleChIPseqRawDataRep1Gm12878Yy1	Gm12878	YY1
wgEncodeYaleChIPseqRawDataRep2Gm12878Yy1	Gm12878	YY1
wgEncodeYaleChIPseqRawDataRep1K562bYy1	K562	YY1
wgEncodeYaleChIPseqRawDataRep2K562bYy1	K562	YY1
wgEncodeHudsonalphaChipSeqRawDataRep1Gm12878Zbtb33	Gm12878	ZBTB33
wgEncodeHudsonalphaChipSeqRawDataRep2Gm12878Zbtb33	Gm12878	ZBTB33
wgEncodeYaleChIPseqRawDataRep1K562bZnf263	K562	ZNF263
wgEncodeYaleChIPseqRawDataRep2K562bZnf263	K562	ZNF263
wgEncodeYaleChIPseqRawDataRep1K562bZnf274	K562	ZNF274
wgEncodeYaleChIPseqRawDataRep2K562bZnf274	K562	ZNF274
wgEncodeSydhTfbsGm12878Znf274UcdRawDataRep1	Gm12878	ZNF274
(continued...)		

Table S1: List of all TF ChIP-seq experiments analyzed in the study.

File name	Cell	TF
wgEncodeYaleChIPseqRawDataRep1Gm12878Zzz3	Gm12878	ZZZ3
wgEncodeYaleChIPseqRawDataRep2Gm12878Zzz3	Gm12878	ZZZ3

Table S2: List of all histone modification ChIP-seq experiments analyzed in the study.

Replicate	Cell	File name
Rep1	Gm12878	wgEncodeBroadChipSeqRawDataRep1Gm12878Control
Rep1	Gm12878	wgEncodeBroadChipSeqRawDataRep1Gm12878Ctcf
Rep1	Gm12878	wgEncodeBroadChipSeqRawDataRep1Gm12878H3k27ac
Rep1	Gm12878	wgEncodeBroadChipSeqRawDataRep1Gm12878H3k27me3
Rep1	Gm12878	wgEncodeBroadChipSeqRawDataRep1Gm12878H3k36me3
Rep1	Gm12878	wgEncodeBroadChipSeqRawDataRep1Gm12878H3k4me1
Rep1	Gm12878	wgEncodeBroadChipSeqRawDataRep1Gm12878H3k4me2
Rep1	Gm12878	wgEncodeBroadChipSeqRawDataRep1Gm12878H3k4me3
Rep1	Gm12878	wgEncodeBroadChipSeqRawDataRep1Gm12878H3k9ac
Rep1	Gm12878	wgEncodeBroadChipSeqRawDataRep1Gm12878H4k20me1
Rep2	Gm12878	wgEncodeBroadChipSeqRawDataRep2Gm12878Control
Rep2	Gm12878	wgEncodeBroadChipSeqRawDataRep2Gm12878Ctcf
Rep2	Gm12878	wgEncodeBroadChipSeqRawDataRep2Gm12878H3k27ac
Rep2	Gm12878	wgEncodeBroadChipSeqRawDataRep2Gm12878H3k27me3
Rep2	Gm12878	wgEncodeBroadChipSeqRawDataRep2Gm12878H3k36me3
Rep2	Gm12878	wgEncodeBroadChipSeqRawDataRep2Gm12878H3k4me1
Rep2	Gm12878	wgEncodeBroadChipSeqRawDataRep2Gm12878H3k4me2
Rep2	Gm12878	wgEncodeBroadChipSeqRawDataRep2Gm12878H3k4me3
Rep2	Gm12878	wgEncodeBroadChipSeqRawDataRep2Gm12878H3k9ac
Rep2	Gm12878	wgEncodeBroadChipSeqRawDataRep2Gm12878H4k20me1
Rep1	Gm12878	wgEncodeUwDnaseSeqRawDataRep1Gm12878
Rep2	Gm12878	wgEncodeUwDnaseSeqRawDataRep2Gm12878
Rep1	K562	wgEncodeBroadChipSeqRawDataRep1K562Control
Rep1	K562	wgEncodeBroadChipSeqRawDataRep1K562Ctcf
Rep1	K562	wgEncodeBroadChipSeqRawDataRep1K562H3k27ac
Rep1	K562	wgEncodeBroadChipSeqRawDataRep1K562H3k27me3
Rep1	K562	wgEncodeBroadChipSeqRawDataRep1K562H3k36me3
Rep1	K562	wgEncodeBroadChipSeqRawDataRep1K562H3k4me1
Rep1	K562	wgEncodeBroadChipSeqRawDataRep1K562H3k4me2
Rep1	K562	wgEncodeBroadChipSeqRawDataRep1K562H3k4me3
Rep1	K562	wgEncodeBroadChipSeqRawDataRep1K562H3k9ac
Rep1	K562	wgEncodeBroadChipSeqRawDataRep1K562H3k9me1
Rep1	K562	wgEncodeBroadChipSeqRawDataRep1K562H4k20me1
Rep1	K562	wgEncodeBroadChipSeqRawDataRep1K562Pol2b
Rep2	K562	wgEncodeBroadChipSeqRawDataRep2K562Control
Rep2	K562	wgEncodeBroadChipSeqRawDataRep2K562Ctcf
(continued...)		

Table S2: List of all histone modification ChIP-seq experiments analyzed in the study.

Replicate	Cell	File name
Rep2	K562	wgEncodeBroadChipSeqRawDataRep2K562H3k27ac
Rep2	K562	wgEncodeBroadChipSeqRawDataRep2K562H3k27me3
Rep2	K562	wgEncodeBroadChipSeqRawDataRep2K562H3k36me3
Rep2	K562	wgEncodeBroadChipSeqRawDataRep2K562H3k4me1
Rep2	K562	wgEncodeBroadChipSeqRawDataRep2K562H3k4me2
Rep2	K562	wgEncodeBroadChipSeqRawDataRep2K562H3k4me3
Rep2	K562	wgEncodeBroadChipSeqRawDataRep2K562H3k9ac
Rep2	K562	wgEncodeBroadChipSeqRawDataRep2K562H3k9me1
Rep2	K562	wgEncodeBroadChipSeqRawDataRep2K562H4k20me1
Rep2	K562	wgEncodeBroadChipSeqRawDataRep2K562Pol2b
Rep1	K562	wgEncodeUwDnaseSeqRawDataRep1K562
Rep2	K562	wgEncodeUwDnaseSeqRawDataRep2K562

Table S3: AUCs for binding prediction experiments using various DNA sequence signal predictors. For each TF, 1000 peaks/flanks are used in the training and test sets.

TF	di-mismatch SVM	cERMIT	MDscan	WeederK1
AP2A1	0.9363	0.8815	0.9151	0.8388
TFAP2C	0.93335	0.9082	0.91815	0.83795
ATF3	0.77215	0.72645	0.72865	0.7195
SMARCC1	0.75785	0.7256	0.74995	0.7195
SMARCC2	0.8349	0.6588	0.7235	0.7962
BATF	0.9356	0.8615	0.9284	0.87395
BCL11A	0.8979	0.7131	0.7069	0.8182
BCL3	0.78385	0.71705	0.676	0.7176
BDP1	0.67286	0.58748	0.62122	0.53528
BRF1	0.6481	0.598625	0.6079	0.545225
BRF2	0.64485	0.60405	0.59095	0.52835
BRG1	0.61955	0.593775	0.595125	0.580825
FOS	0.92425	0.8599125	0.867825	0.86965
JUN	0.9451	0.882616667	0.884883333	0.873716667
MYC	0.8965	0.85665	0.83755	0.8332
CTCF	0.98085	0.8778	0.9821	0.91345
E2F1	0.9018	0.783	0.804	0.8184
E2F4	0.87215	0.807375	0.7934	0.776925
E2F6	0.86095	0.78145	0.76685	0.766575
EGR1	0.900066667	0.847166667	0.812066667	0.7928
GABPA	0.9517	0.926716667	0.9203	0.918283333
GATA1	0.91615	0.8755	0.86625	0.8387
GATA2	0.9265	0.84925	0.8662	0.8366
GTF2B	0.677	0.5972	0.5853	0.5941
HA-E2F1	0.92985	0.80705	0.84275	0.85335
(continued...)				

Table S3: AUCs for binding prediction experiments using various DNA sequence signal predictors. For each TF, 1000 peaks/flanks are used in the training and test sets.

TF	di-mismatch SVM	cERMIT	MDscan	WeederK1
HEY1	0.71905	0.609	0.5812	0.58545
SMARCB1	0.61825	0.57805	0.582575	0.580775
IRF4	0.928	0.77905	0.7569	0.8477
JUND	0.935233333	0.8847	0.8823	0.889383333
MAX	0.9361	0.901883333	0.89825	0.885666667
RDBP	0.58025	0.55865	0.5772	0.55355
NFE2	0.9738	0.8992	0.6717	0.88255
NFKB	0.8996	0.8319125	0.86445	0.84575
NYFA	0.9517	0.93705	0.92785	0.92975
NYFB	0.93685	0.8942	0.92385	0.9117
NRF1	0.9192	0.88735	0.909	0.9108
REST	0.9941	0.949225	0.949225	0.954425
EP300	0.74685	0.61415	0.5669	0.68495
PAX5	0.84955	0.73905	0.79855	0.7723
PBX3	0.87555	0.7967	0.8269	0.8263
POLR2	0.718642857	0.625785714	0.62557619	0.616347619
POLR3	0.6006	0.5475	0.546425	0.524875
POU2F2	0.8009	0.73875	0.7316	0.7155
SPI1	0.99265	0.9629	0.98815	0.9396
RAD21	0.97405	0.944075	0.983325	0.895175
POLR3A	0.761675	0.6357	0.60995	0.626575
SETDB1	0.6629	0.613275	0.5997	0.590375
SIN3A	0.785475	0.6631	0.65105	0.687275
SIRT6	0.8146	0.67305	0.6548	0.741
SIX5	0.8946	0.8283	0.8415	0.82425
SP1	0.9425	0.916166667	0.935166667	0.906433333
SRF	0.838125	0.68805	0.6807	0.722425
STAT1	0.88095	0.84825	0.87135	0.81765
TAF1	0.7363	0.643583333	0.64105	0.641983333
TAF7	0.8278	0.704	0.7196	0.7123
TBP	0.7063	0.61265	0.5946	0.57625
TCF12	0.92765	0.8821	0.91265	0.86605
GTF3C2	0.75706	0.6788	0.67422	0.71484
NR2C2	0.804616667	0.653733333	0.65195	0.7019
USF1	0.9796	0.979275	0.9798	0.957025
WRNIP1	0.4709	0.6553	0.5969	0.4866
XRCC4	0.5964	0.54885	0.53065	0.5213
YY1	0.898375	0.863025	0.8767	0.853775
ZBTB33	0.86645	0.807	0.811	0.80905
ZNF263	0.8311	0.7669	0.76935	0.764
ZNF274	0.61425	0.59935	0.60555	0.57175
(continued...)				

Table S3: AUCs for binding prediction experiments using various DNA sequence signal predictors. For each TF, 1000 peaks/flanks are used in the training and test sets.

TF	di-mismatch SVM	cERMIT	MDscan	WeederK1
ZZZ3	0.82825	0.6844	0.76895	0.7336

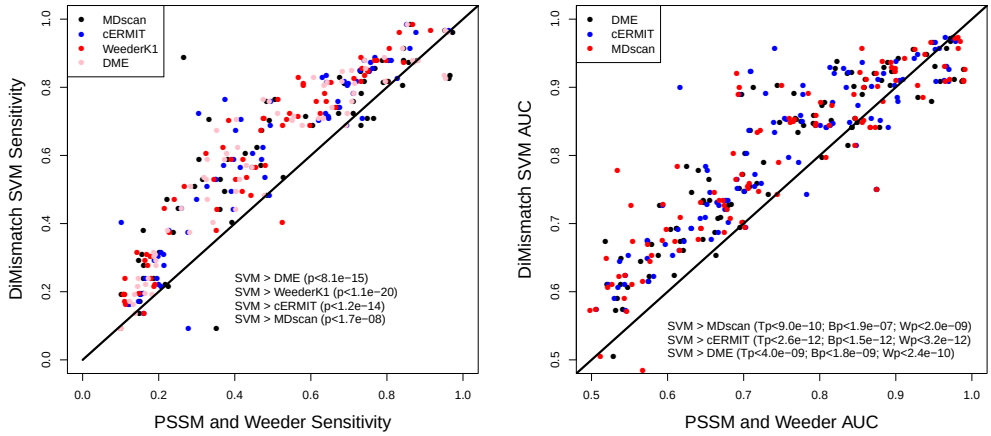


Figure S1: **Additional accuracy measures are consistent with AUC.** We examined additional accuracy measures, including sensitivity at a 0.1 false positive rate (left) and evaluating on the top 5000 peaks or number of peaks found with 20% FDR cutoff by SPP (right). We found that both measures support the increase in accuracy we see for the SVM approach.

Table S4: AUCs for binding prediction experiments using DNase, kmer-SVM, or a combination of DNase and kmer-SVM. For each TF, 5000 (or up to 20% FDR) peaks/flanks are used in the training/test sets.

TF	DNase	Kmer-SVM	Combined
ATF3	0.847	0.753	0.855
BATF	0.97	0.931	0.988
BCL11A	0.977	0.894	0.985
BCL3	0.846	0.769	0.882
BDP1	0.604	0.638	0.645
BRF1	0.546	0.654	0.604
BRF2	0.717	0.641	0.743
BRG1	0.748	0.643	0.743
FOS	0.921	0.826	0.93
JUN	0.98	0.903	0.991
MYC	0.922	0.838	0.952
CTCF	0.986	0.909	0.992
E2F4	0.914	0.857	0.945
E2F6	0.844	0.85	0.91
(continued...)			

Table S4: AUCs for binding prediction experiments using DNase, kmer-SVM, or a combination of DNase and kmer-SVM. For each TF, 5000 (or up to 20% FDR) peaks/flanks are used in the training/test sets.

TF	DNase	Kmer-SVM	Combined
EBF1	0.929	0.867	0.965
EGR1	0.827	0.852	0.894
ELF1	0.947	0.892	0.971
GABPA	0.913	0.912	0.957
GATA1	0.909	0.866	0.932
GATA2	0.933	0.879	0.948
GTF2B	0.829	0.707	0.829
HEY1	0.848	0.746	0.876
SMARCB1	0.731	0.648	0.748
Input-Control	0.642	0.524	0.595
IRF3	0.912	0.826	0.938
IRF4	0.984	0.905	0.99
JUND	0.847	0.828	0.887
MAX	0.895	0.896	0.955
MEF2C	0.896	0.728	0.909
RDBP	0.71	0.75	0.775
NFE2	0.933	0.933	0.97
NFE2	0.97	0.907	0.984
NFKB	0.933	0.872	0.967
NYFA	0.936	0.92	0.978
NYFB	0.917	0.89	0.949
NRF1	0.932	0.959	0.978
REST	0.704	0.928	0.875
EP300	0.878	0.729	0.879
PAX5	0.935	0.841	0.957
PBX3	0.878	0.776	0.896
POLR2	0.827	0.721	0.849
POLR3	0.574	0.595	0.601
POU2F2	0.898	0.771	0.917
SPI1	0.929	0.91	0.97
RAD21	0.994	0.906	0.996
RFX5	0.927	0.903	0.963
POLR3A	0.684	0.665	0.715
SETDB1	0.669	0.671	0.712
SIN3A	0.779	0.803	0.833
SIRT6	0.864	0.747	0.872
SIX5	0.959	0.883	0.972
SMC3	0.988	0.887	0.991
SP1	0.921	0.9	0.973
SRF	0.801	0.808	0.856
(continued...)			

Table S4: AUCs for binding prediction experiments using DNase, kmer-SVM, or a combination of DNase and kmer-SVM. For each TF, 5000 (or up to 20% FDR) peaks/flanks are used in the training/test sets.

TF	DNase	Kmer-SVM	Combined
STAT1	0.744	0.447	0.654
STAT3	0.814	0.568	0.819
TAF1	0.832	0.781	0.879
TAF7	0.854	0.878	0.909
TBP	0.896	0.665	0.895
TCF12	0.947	0.863	0.975
GTF3C2	0.701	0.719	0.777
NR2C2	0.805	0.803	0.847
USF1	0.887	0.919	0.956
USF2	0.936	0.913	0.975
WRNIP1	0.768	0.581	0.736
XRCC4	0.58	0.593	0.613
YY1	0.832	0.868	0.909
ZBTB33	0.921	0.823	0.953
ZNF143	0.955	0.798	0.968
ZNF263	0.664	0.75	0.765
ZNF274	0.576	0.717	0.659
ZZZ3	0.678	0.924	0.881

Table S5: AUCs for binding prediction experiments using DNase, kmer-SVM, or a combination of DNase and kmer-SVM. Accuracy values are given for training/testing on the same cell type and when trained on one cell line and tested on the other (Trans).

TF	DNase	Kmer-SVM	Combined	DNase (Trans)	Kmer-SVM (Trans)	Combined (Trans)
FOS	0.921	0.826	0.93	0.92	0.826	0.915
EGR1	0.827	0.852	0.894	0.816	0.83	0.879
GABPA	0.913	0.912	0.957	0.915	0.92	0.958
JUND	0.847	0.828	0.887	0.849	0.659	0.813
MAX	0.895	0.896	0.955	0.894	0.898	0.958
REST	0.704	0.928	0.875	0.695	0.937	0.919
POLR2	0.827	0.721	0.849	0.802	0.648	0.771
POLR3	0.574	0.595	0.601	0.582	0.57	0.298
SIN3A	0.779	0.803	0.833	0.775	0.769	0.827
SP1	0.921	0.9	0.973	0.895	0.955	0.98
SRF	0.801	0.808	0.856	0.802	0.77	0.852
TAF1	0.832	0.781	0.879	0.847	0.766	0.867
NR2C2	0.805	0.803	0.847	0.801	0.633	0.792
USF1	0.887	0.919	0.956	0.888	0.943	0.973
YY1	0.832	0.868	0.909	0.826	0.913	0.932

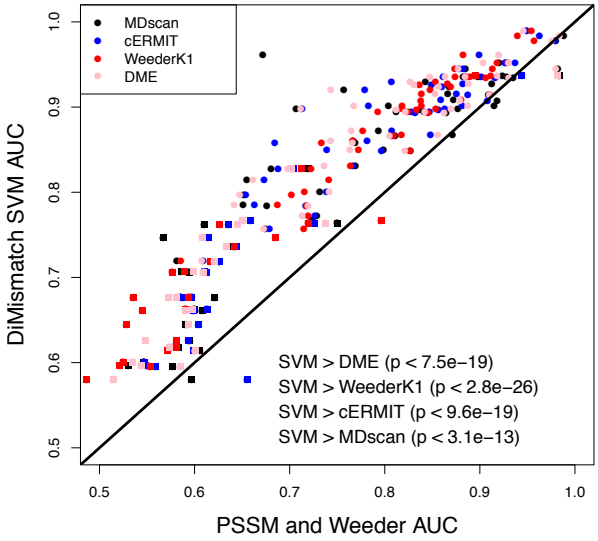
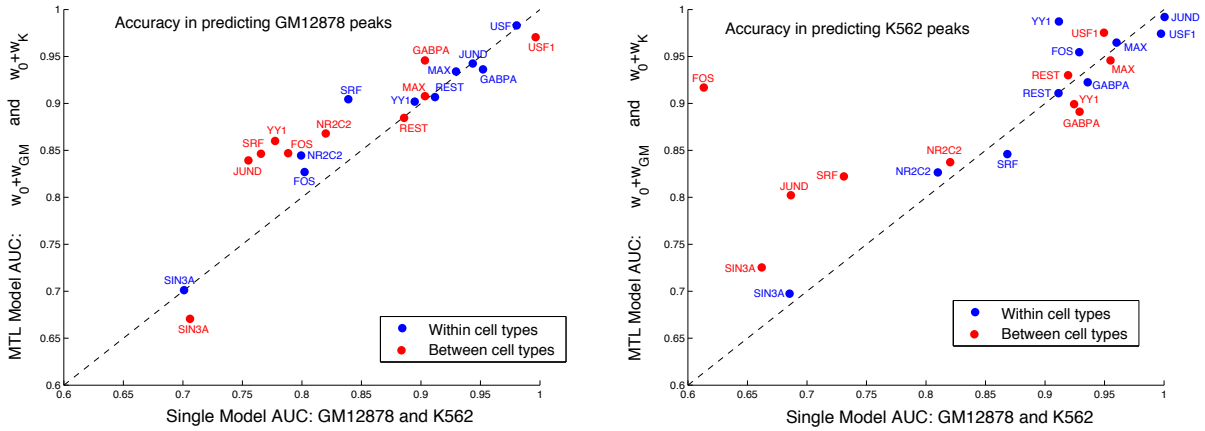


Figure S2: **SVM models better capture non-sequence factors.** Accuracy on sequence specific factors shown by circles, non-sequence specific or indirectly binding TFs shown as squares. Significance is determined on all TFs and determined by signed rank test.

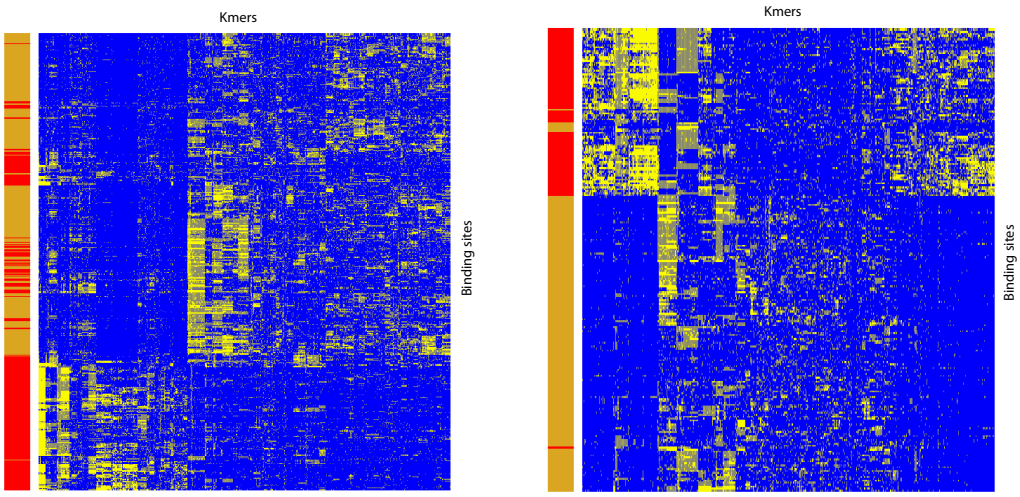
TF	Total Peaks	Cell-Specific Peaks	Cell-Exclusive Peaks (RPM < 2)	Cell-Exclusive Peaks (RPM < 1)
TAF1	3630	241 (0.066)	106 (0.029)	52 (0.014)
MAX	4866	1558 (0.320)	1212 (0.249)	894 (0.183)
FOS	4350	285 (0.066)	183 (0.042)	115 (0.026)
JUN	5450	2243 (0.412)	2064 (0.379)	1824 (0.334)
JUND	3946	1260 (0.319)	1189 (0.301)	983 (0.249)
YY1	3705	1440 (0.389)	767 (0.207)	478 (0.129)
REST	4967	1792 (0.361)	594 (0.119)	370 (0.074)
USF1	5547	1574 (0.284)	693 (0.125)	405 (0.073)
GABPA	5084	1077 (0.212)	473 (0.093)	305 (0.059)
SRF	2744	722 (0.263)	488 (0.178)	346 (0.126)

Table S6: **Cell-type specific peaks.** Percentage of top 5000 peaks from both cell lines that are supported by differential read counts in GM12878 versus K562 ($p < 0.01$) based on a replicate-to-replicate noise model.



(a) (b)

Figure S3: **Multi-task learning improves over training models independently on cell types.** (a) A comparison of performance SVM sequence models that are independently trained on cell-type exclusive binding sites from GM12878 or K562 (x-axis) versus cell-type specific SVM models that are jointly trained on cell-type exclusive sites from both GM12878 and K562 using multi-task learning (y-axis). The test set is also sampled from GM12878-exclusive binding sites and excludes training sequences. Multi-task learning yields both a GM12878-specific model (model vector $\mathbf{w}_0 + \mathbf{w}_{\text{GM}}$) and a K562-specific model (model vector $\mathbf{w}_0 + \mathbf{w}_{\text{K}}$). The GM12878-specific model from multi-task learning outperforms the independently trained model for GM12878 on GM12878-exclusive sites. Interestingly, the K562-specific multi-task model also outperforms the independently trained K562 models on these sites. (b) The same analysis but comparing multi-task models to independently trained model independently on a test set of K562-exclusive binding sites.



(a) JUND (b) YY1

Figure S4: **Clustering of k-mer feature vector for cell-type exclusive JUND and YY1 binding sites reveals subtype-specific patterns.** The heatmaps show the k-mer feature vectors used to train our cell-type specific SVM sequence models for (a) JUND and (b) YY1. In these heatmaps, each row is the dinucleotide mismatch k-mer feature vector for a cell-type exclusive binding site. Rows labeled in gold are GM12878-exclusive sites, and rows labeled in red are K562-exclusive sites. Each column represents a k-mer feature. Feature values have been binarized give a clearer visualization, with 0 values in blue and non-zero values in yellow. Note that the k-mers are selected to discriminate GM12878 or K562 binding sites from their flanking regions, not to distinguish the GM12878- and K562-exclusive binding sites from each other. Clustering of rows columns reveals blocks of co-occurring k-mers that are strongly enriched in either the GM12878 or K562 sites.

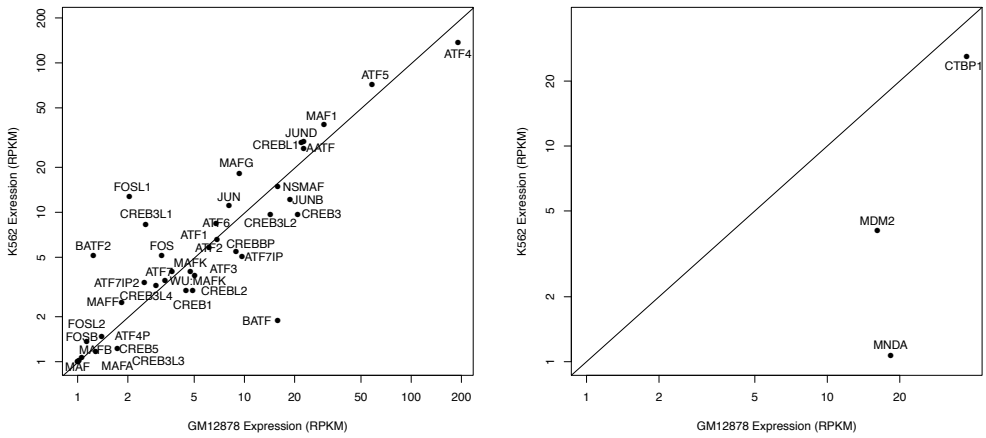
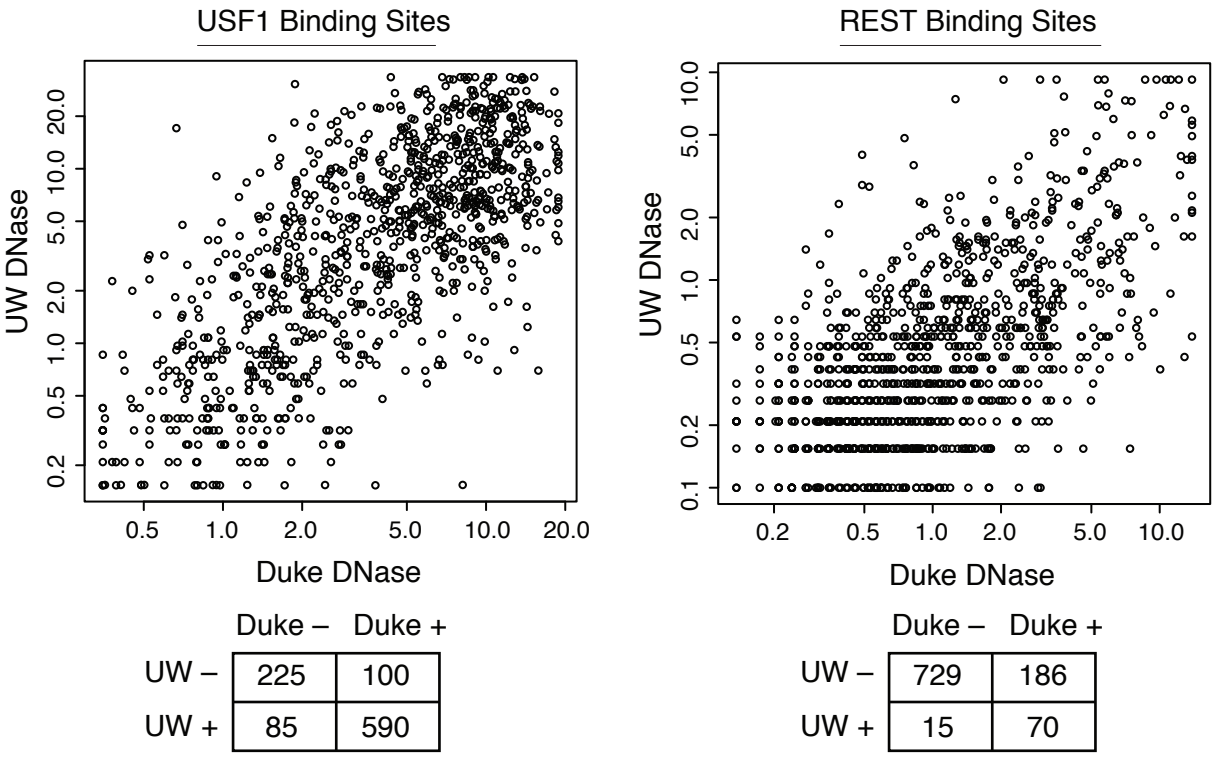


Figure S5: **Components of AP-1 and YY1 complexes display differential expression in GM12878 and K562.** RNA-seq analysis of differential expression levels of various cofactors of the AP-1 (left) and YY1 (right) complexes. BATF and FOS-like proteins in particular are known to form heterodimers that can provide the differential JUND sequence specificity and are differentially expressed. Similarly, cofactors for YY1, such as myeloid cell nuclear differentiation antigen (MNDA), are expressed at a much higher level in GM12878.



	Duke –	Duke +
UW –	225	100
UW +	85	590

	Duke –	Duke +
UW –	729	186
UW +	15	70

Figure S6: **DNase-seq from two sources display subtle differences.** In the main text, we examined DNase-seq from the University of Washington (UW). When we compared to DNase-seq data from Duke, we noticed that data were fairly consistent on both USF1 peaks and REST peaks. Scatterplot data are quantile normalized, and the top/bottom 5% of sites are capped to emphasize the dynamic range of the middle 90%. We noticed that DNase accessible regions showed more overlap for USF1 (an activator) in contrast to REST (a repressor); however it is unclear if this pattern generalizes as most TFs examined act mostly as activators and several TFs act pleiotropically.