Supplementary Material

Conserved and species-specific transcription factor co-binding patterns drive divergent gene regulation in human and mouse

Adam G. Diehl[1], Alan P. Boyle[1,2]*

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
[2]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA

List of Supplementary Items

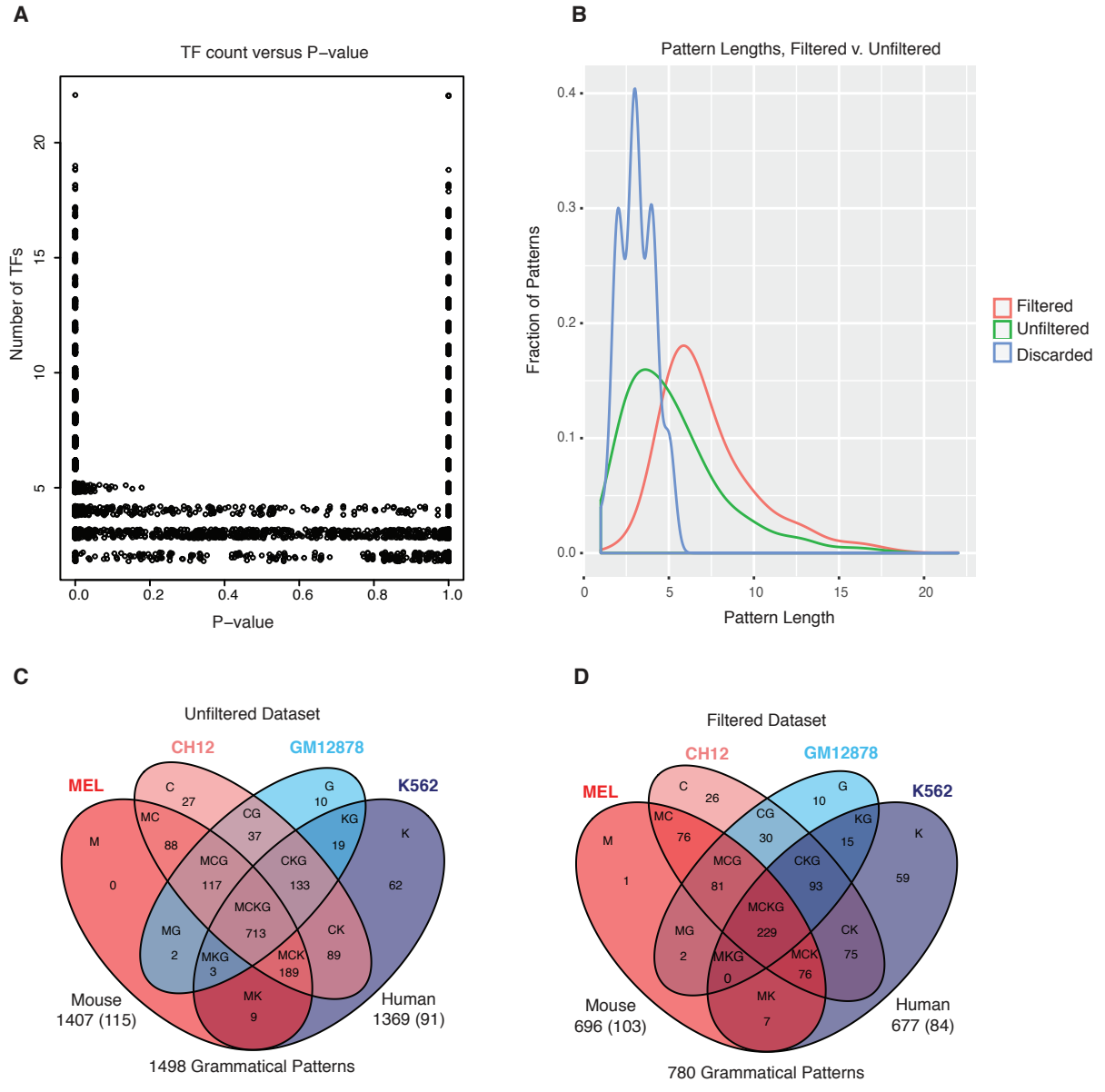

Figure S1. Effects of removing patterns present in 10,000 random data permutations. **(A)** Distribution of p-values relative to grammatical pattern lengths in the unfiltered dataset shows no apparent difference between long and short patterns at either end of the p-value spectrum. **(B)** Length distributions of grammatical patterns in the unfiltered dataset, filtered, and discarded fractions of the dataset. Although the bulk of filtered patterns contained between 2 and 5 factors, removing these patterns caused a mean shift of only 2 units between the unfiltered and filtered datasets, with no effect on maximum and minimum pattern lengths. **(C-D)** Venn diagrams show the segmentation of regulatory space into 15 possible grammatical and positional classes in the unfiltered dataset (C) and filtered dataset(D). The first letter of each cell type was used to construct a class label for each cell in the diagrams. These labels describe the cell-specificity of the corresponding grammatical patterns and positionally-conserved loci. Each segment in the Venn diagrams is labelled with its grammatical class and the number of grammatical patterns assigned to the class. **(C)** The unfiltered dataset contains a total of 1498 grammatical patterns, 86% of which are shared between human and mouse. **(D)** The filtered dataset contains a total of 780 grammatical patterns, 76% of which are shared between human and mouse. Therefore, data filtration did not bias our observations towards increased conservation.

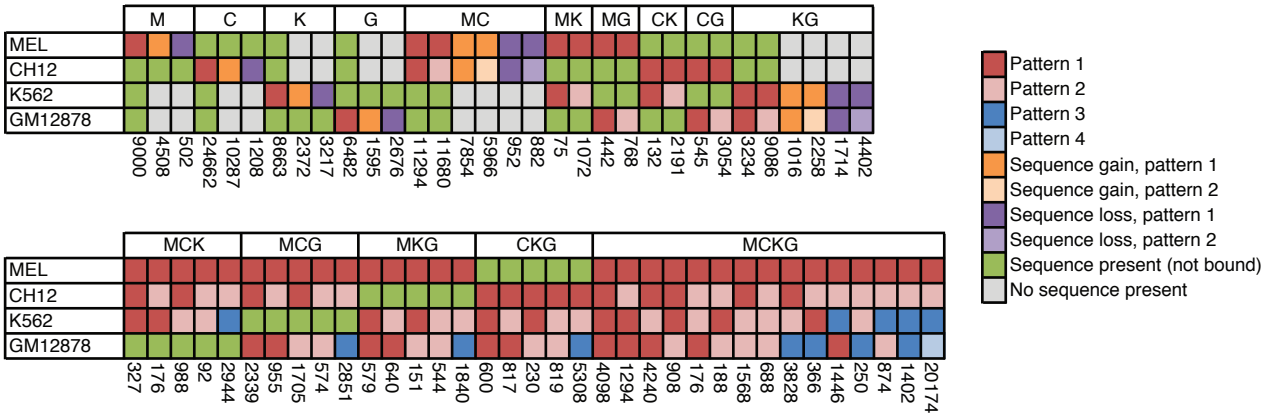

Figure S2. Grammatical pattern distributions within positional classes.
Each column in the matrix represents a combination of matched or mismatched patterns between sequences bound at the same locus. Column headings indicate the positional class represented by a set of columns and the number below each column represents the total number of CRMs within the subset. Cells with matching colors in a column share the same grammatical pattern while cells with mismatched colors are occupied by different grammatical patterns. Green cells indicate sequences which are physically present, but are not occupied in the given cell type. Grey cells indicate sequences which are physically absent in the given species as a result of sequence gain or loss. Orange hues indicate species-specific sequence gains and purple hues indicate sequence losses in the other species, based on phylogenetic maximum parsimony prediction using three outgroup species.

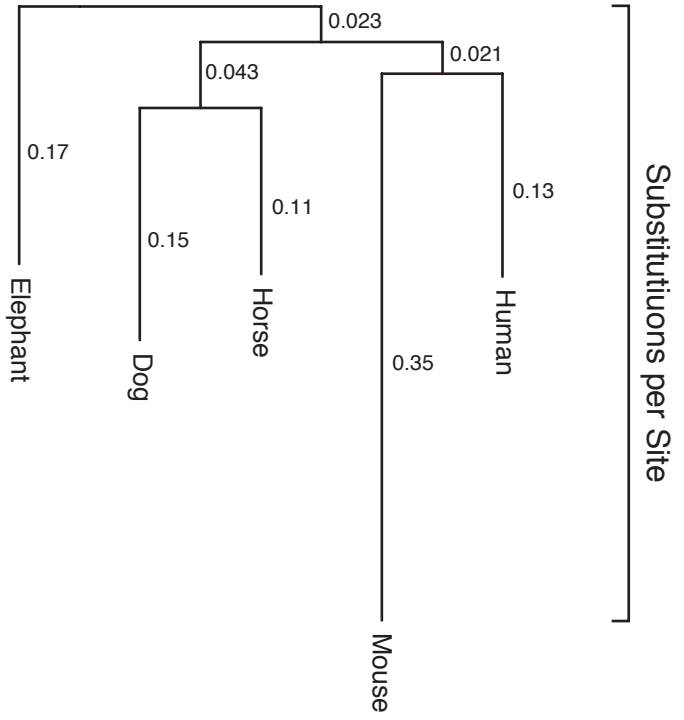

Figure S3. Gain/loss phylogeny.
Phylogenetic tree describing evolutionary relationships between human, mouse and the outgroup species horse, dog and elephant, used in assigning non-orthologous CRMs as species-specific gains or losses.

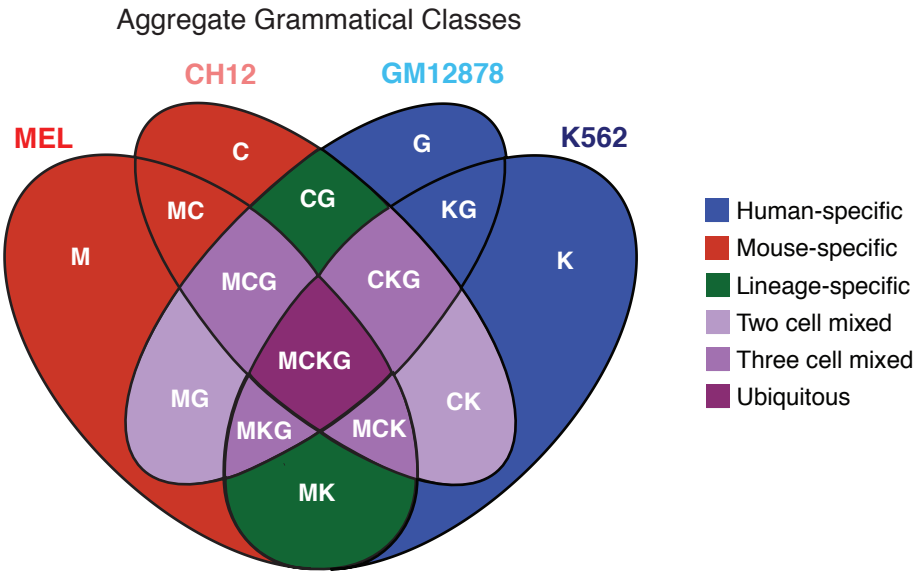

Figure S4. Aggregate grammatical classes explained.
Grammatical classes were aggregated into six species-specific and tissue-specific aggregate grammatical classes to facilitate further analysis of functional specificity.

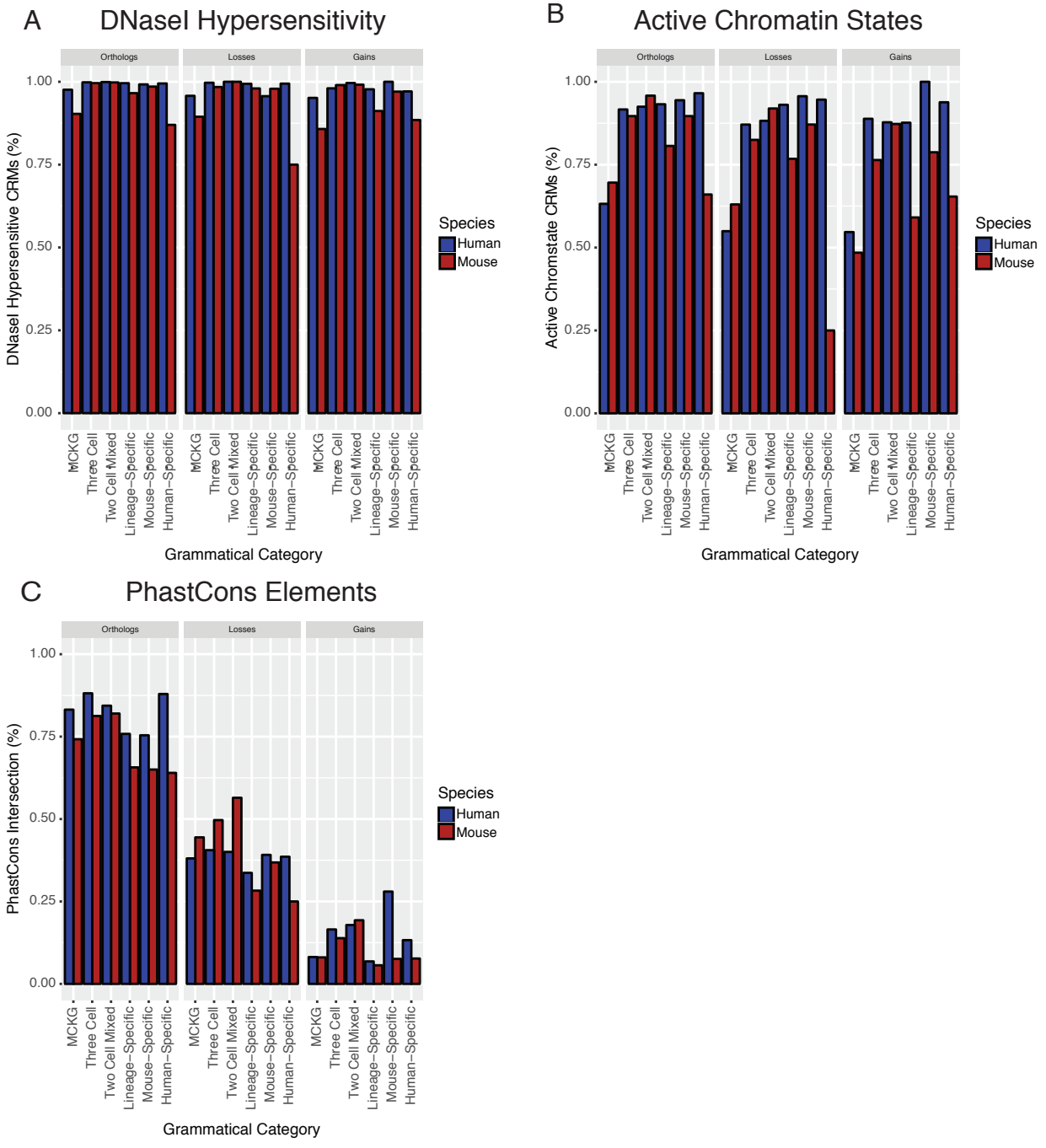

Figure S5. Frequency histograms of functional annotations associated with regulatory function in six aggregate grammatical classes after excluding the two largest cohesin-related grammatical patterns from the MCKG class.
(A) Intersection with DNaseI Hypersensitive Sites **(B)** Intersection with ChromHmm active chromatin states. **(C)** Intersection with PhastCons elements.

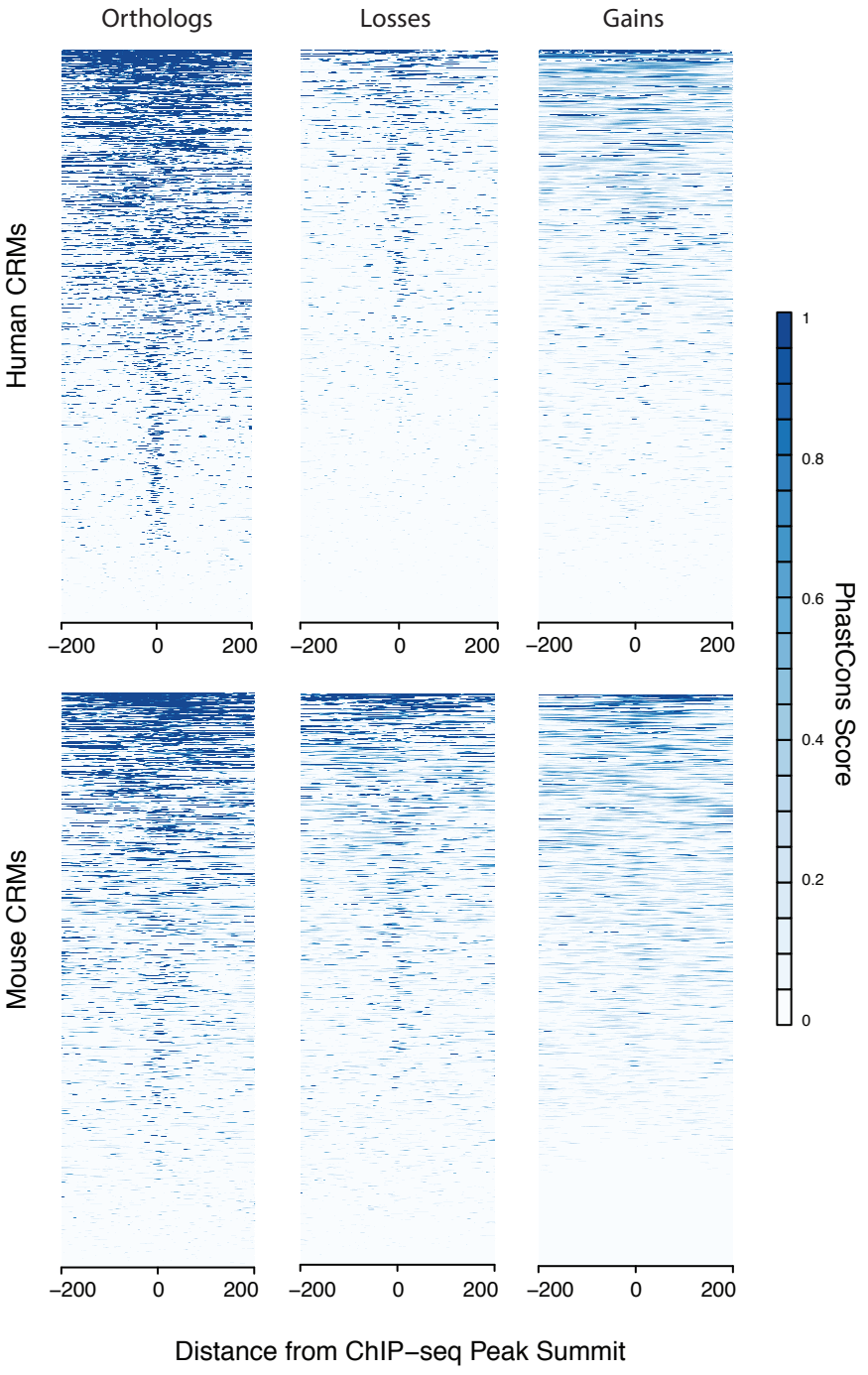

Figure S6. Conserved PhastCons scores cluster near ChIP-seq peak summits in all fractions of the dataset.

PhastCons scores within 400bp windows centered over ChIP-seq peak summits from 1000 randomly-selected CRMs classified as orthologs, losses, and gains in Human and Mouse. Clustering of scores approaching 1 centered around ChIP-seq peak summits suggests significant contributions of occupied TFBS to overall conservation in many CRMs. Scores approaching 1 in more distant foci may correspond with additional TFBSs for which we lack ChIP-seq data, TFBSs occupied only in other cell types, or other classes of functional sequence not annotated in the current study. Longer stretches of high PhastCons scores, concentrated among the top rows of all groups, likely reflect the "smoothing" effect of the PhastCons hidden Markov model over stretches of DNA containing multiple conserved features in close proximity.

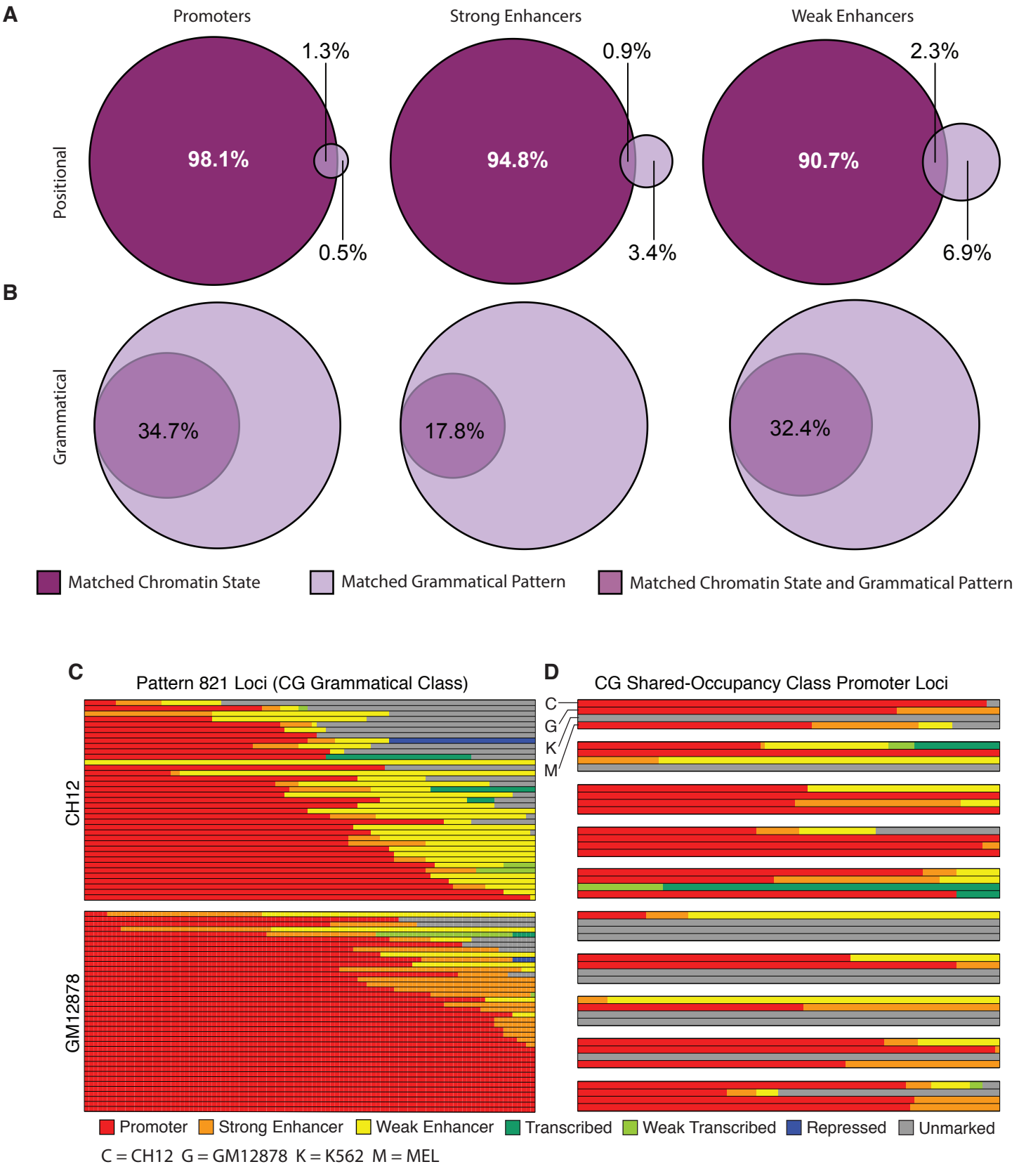

Figure S7. Grammatical patterns predict matched chromatin states more accurately than positional conservation of chromatin states predicts matched grammatical patterns.
(A) Fractions of positionally-conserved locus pairs with matched Promoter, Strong Enhancer, or Weak Enhancer chromatin states and/or grammatical patterns. **(B)** Fractions of locus pairs within grammatical patterns where the underlying Promoter, Strong Enhancer, or Weak Enhancer chromatin state

also matches. **(C)** Chromatin states within all 77 CRMs CRMs in grammatical pattern 821 (CG grammatical class (M and K). (D) Ten randomly-selected orthologous loci representing positional class CG and containing the promoter state in one or both of C and G. Each block represents a different genomic locus, with individual rows representing the chromatin states observed in the cell type indicated.

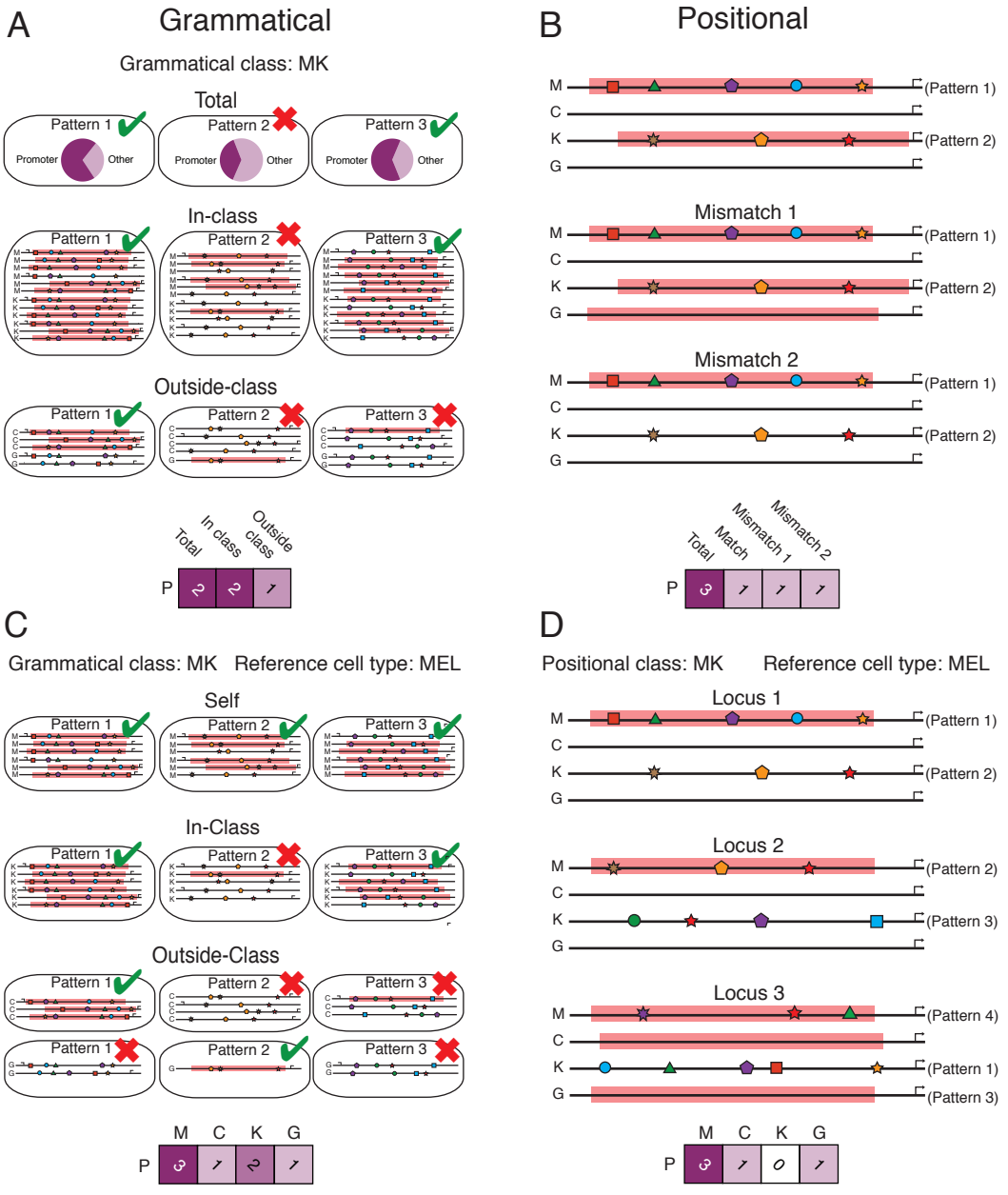

Figure S8. Graphical definition of counting conventions used in chromatin state heatmaps. In all panels, a theoretical set of loci or patterns are given with associated data for the promoter chromatin state, and the accompanying heatmap shows the result from the applicable counting procedure. **(A)** For Fig 4C, counts are aggregated over all grammatical patterns. All grammatical patterns in which >= 50% of CRMs carry a given state are counted in the total column. Among these patterns, In-class: >= 50% of CRMs in cells belonging to the grammatical class carry the state. Outside class: >= 50% of modules from nonmember cells carry the state. **(B)** For Fig 4D, counts are pooled over all positionally-conserved loci. All loci carrying a given state in >= 1 occupied cell are counted toward the total column. Match: the cell-specificity of chromatin states matches that expected based on the positional class. Mismatch 1: 1 or more cells not included in the positional class carry

the given state. Mismatch 2: 1 or more cells within the positonal class lack the given state. **(C)** For sup fig 5A, counts represent the number of times a chromatin mark was observed in >= 50% of peaks form a given pattern within a grammatical class in one cell type given that it was observed in >= 50% of peaks from the pattern in a reference cell type. **(D)** For sup fig 5B, counts represent the number of observations in which a chromatin mark was observed at a positional locus in one cell given that it

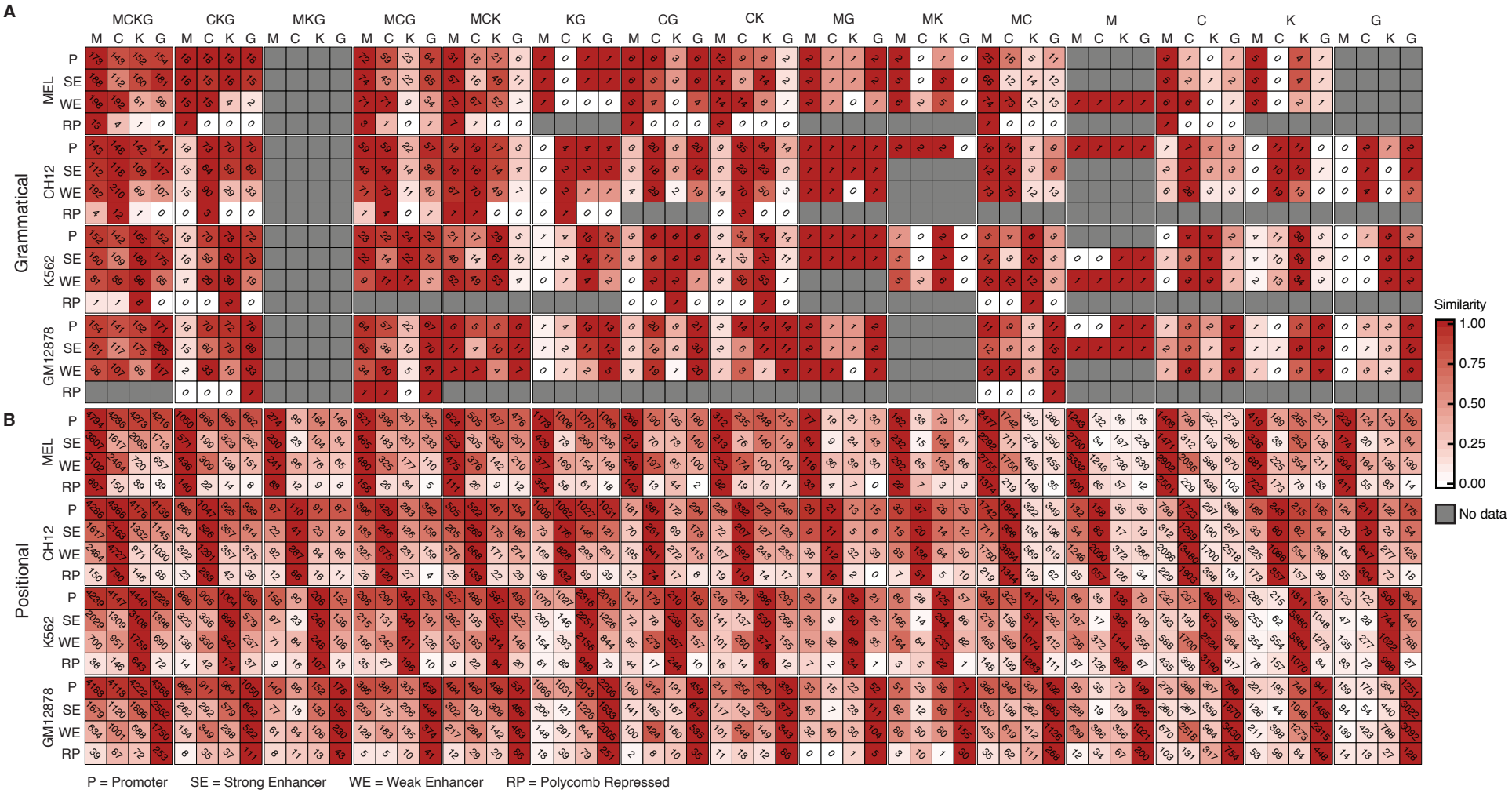

Figure S9. Chromatin state overlaps in individual cell types by grammatical and positional class.
Heat maps describe the degree of conservation of ChromHMM chromatin states between loci in positional classes and patterns in grammatical classes representing a non-collapsed view of Fig. 4. Shading densities within each cell represent the fraction of loci or patterns that carry the given chromatin state marks. **(A)** Within grammatical classes, the frequency of overlap for each cell type with a matching module annotation in at least 50% of the loci from all other cell types. This is analogous to “inside class”, on a per cell basis, from Fig. 4C. **(B)** Within positional classes, the frequency of overlap for each cell type with a matching module annotation in at the same loci from all other cell types. This is analogous to “match” on a per cell basis from Fig. 4D.

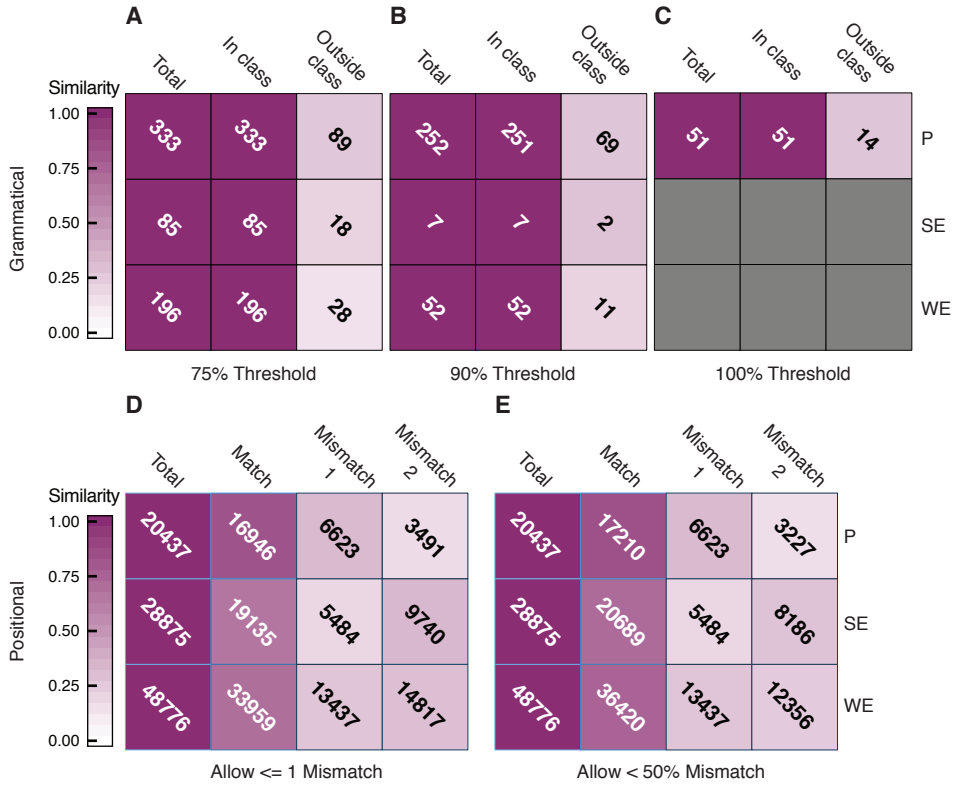

Figure S10. Cell-specificity of chromatin states is not affected by inclusion thresholds. **(A-C)** Thresholds for grammatical patterns are arranged in order of increasing stringency. **(A)** Cell-specificity of chromatin states for grammatical patterns, requiring >= 75% of CRMs within a pattern include a given state. **(B)** Cell-specificity of chromatin states for grammatical patterns, requiring >= 90% of CRMs within a pattern include a given state. **(C)** Cell-specificity of chromatin states for grammatical patterns, requiring 100% of CRMs within a pattern include a given state. **(D and E)** Thresholds for positional classes are arranged in order of decreasing stringency. **(D)** Cell-specificity of chromatin states for positional classes, allowing up to 1 mismatch before a locus is counted toward either mismatch category. **(E)** Cell-specificity of chromatin states for positional classes, allowing 1 mismatch for loci occupied in 3 cells and up to 2 mismatches in loci occupied in 4 cells, before a locus is counted toward either mismatch categories.

Table S1.
Identification of transcription factor binding datasets used in this project. All data were obtained from the ENCODE project through encodeproject.org. Accession numbers are provided along with organism, cell, and other relevant information.

Table S2.
Identification of chromHMM annotations used in this project. All data were obtained from the ENCODE project through encodeproject.org. Accession numbers are provided along with organism, cell, and other relevant information.

Table S3.
Identification of DNase-seq datasets used in this project. All data were obtained from the ENCODE project through encodeproject.org. Accession numbers are provided along with organism, cell, and other relevant information.

Table S4.
Identification of RNA-seq datasetsused in this project. All data were obtained from the ENCODE project through encodeproject.org. Accession numbers are provided along with organism, cell, and other relevant information.

Table S5.
Orthology statistics and PhastCons element intersections for human and mouse CRMs and background sequences. Numbers of CRMs/background sequences, percent sequences mappable by bnMapper, and percent of sequences containing <= 1 phastCons element are given for mouse and human sequences for the total dataset, orthologous CRMs, unmapped CRMs (gains and losses combined), species-specific gains and species-specific losses. Odds ratios and p-values were calculated using individual Fisher's Exact tests against matched background sequences. P-values were corrected for multiple testing with the holm method.

Table S6.
Percentage of CRMs and background sequences overlapping DNaseI hypersensitive (DHS) sites. The number of human and mouse CRMs in each of six grammatical categories are given along with percentages of CRMs and background sequences within each category overlapping a DHS site(s), for the total dataset, orthologous CRMs, unmapped CRMs (gains and losses combined), species-specific gains and species-specific losses. Odds ratios and p-values were calculated using individual Fisher's Exact tests against matched background sequences. P-values were corrected for multiple testing with the holm method.

Table S7.
Percentage of CRMs and background sequences marked with active chromatin states from ChromHMM [27]. The number of human and mouse CRMs in each of six aggregate grammatical categories are given along with percentages of CRMs and background sequences within each category overlapping regions of the genome assigned to active states (states 1-5) by ChromHMM (see table S1), for the total dataset, orthologous CRMs, unmapped CRMs (gains and losses combined), species-specific gains and species-specific losses. Odds ratios and p-values were calculated using individual Fisher's Exact tests against matched background sequences. P-values were corrected for multiple testing with the holm method.

Table S8.
GWAS Ontology terms and associated functional classifications.

Table S9.
GWAS Ontology term enrichments found within human CRMs pooled across all grammatical patterns. For each term, observed counts were collected by counting the occurrences of each term among CRM-associated GWAS and GWAS-linked SNPs within the human dataset to compute the expected binomial frequency. Each term was counted only once per CRM in cases where a CRM contained multiple SNPs annotated with the same term. All terms in the total dataset with uncorrected p-values <= 0.05 are reported in the table but only those passing a 0.05 threshold after multiple testing correction by the FDR method were retained for further analysis. Immune terms are presented in red text.

Table S10.
GWAS Ontology term enrichments by aggregate grammatical class. CRMs were separated by aggregate grammatical class and binomial enrichment tests for each GWAS Ontology term was performed following the same procedures as for the pooled set. Terms presented in blue text were found only in the analysis of aggregate grammatical classes, but not in the pooled dataset. Immune terms are presented in red text.

Table S11.
GWAS Ontology term enrichments by individual grammatical class. CRMs were separated into individual grammatical classes and binomial enrichment tests for each GWAS Ontology term was performed following the same procedures as for the pooled set. Terms presented in blue text were found only in the analysis of individual grammatical classes, but not in the pooled dataset. Immune terms are presented in red text.

Table S1

Accession	Species	Cell	TF	Format
ENCSR000ERC	Mouse	CH12	BHLHE40	narrowPeak
ENCSR000DZJ	Human	GM12878	BHLHE40	narrowPeak
ENCSR000EGV	Human	K562	BHLHE40	narrowPeak
ENCSR000ESH	Mouse	MEL	BHLHE40	narrowPeak
ENCSR000EQV	Mouse	CH12	CHD1	narrowPeak
ENCSR000DZE	Human	GM12878	CHD1	narrowPeak
ENCSR000AQD	Human	K562	CHD1	narrowPeak
ENCSR000ESN	Mouse	MEL	CHD1	narrowPeak
ENCSR000ERF	Mouse	CH12	CHD2	narrowPeak
ENCSR000DZR	Human	GM12878	CHD2	narrowPeak
ENCSR000EHD	Human	K562	CHD2	narrowPeak
ENCSR000ETO	Mouse	MEL	CHD2	narrowPeak
ENCSR000ERM	Mouse	CH12	CTCF	narrowPeak
ENCSR000AKB	Human	GM12878	CTCF	narrowPeak
ENCSR000BPJ	Human	K562	CTCF	narrowPeak
ENCSR000ETE	Mouse	MEL	CTCF	narrowPeak
ENCSR000ERU	Mouse	CH12	E2F4	narrowPeak
ENCSR000DYY	Human	GM12878	E2F4	narrowPeak
ENCSR000EWL	Human	K562	E2F4	narrowPeak
ENCSR000ETY	Mouse	MEL	E2F4	narrowPeak
ENCSR293WTN	Mouse	CH12	ELF1	narrowPeak
ENCSR000BMB	Human	GM12878	ELF1	narrowPeak
ENCSR000BMD	Human	K562	ELF1	narrowPeak
ENCSR033OWC	Mouse	MEL	ELF1	narrowPeak
ENCSR000ERI	Mouse	CH12	EP300	narrowPeak
ENCSR000DZD	Human	GM12878	EP300	narrowPeak
ENCSR000EGE	Human	K562	EP300	narrowPeak
ENCSR000ETP	Mouse	MEL	EP300	narrowPeak
ENCSR000ERA	Mouse	CH12	ETS1	narrowPeak
ENCSR000BKA	Human	GM12878	ETS1	narrowPeak
ENCSR000BKQ	Human	K562	ETS1	narrowPeak
ENCSR000ETB	Mouse	MEL	ETS1	narrowPeak
ENCSR000EQS	Mouse	CH12	GABPA	narrowPeak
ENCSR000BGC	Human	GM12878	GABPA	narrowPeak
ENCSR000BLO	Human	K562	GABPA	narrowPeak
ENCSR000ESK	Mouse	MEL	GABPA	narrowPeak
ENCSR000ERR	Mouse	CH12	JUND	narrowPeak
ENCSR000EYV	Human	GM12878	JUND	narrowPeak
ENCSR000EGN	Human	K562	JUND	narrowPeak
ENCSR000ETZ	Mouse	MEL	JUND	narrowPeak
ENCSR000ERB	Mouse	CH12	MAFK	narrowPeak
ENCSR000DYV	Human	GM12878	MAFK	narrowPeak
ENCSR000EGX	Human	K562	MAFK	narrowPeak
ENCSR000ETK	Mouse	MEL	MAFK	narrowPeak
ENCSR000ERL	Mouse	CH12	MAX	narrowPeak
ENCSR000DZF	Human	GM12878	MAX	narrowPeak
ENCSR000EFV	Human	K562	MAX	narrowPeak
ENCSR000ETX	Mouse	MEL	MAX	narrowPeak
ENCSR000EQT	Mouse	CH12	MAZ	narrowPeak
ENCSR000DZA	Human	GM12878	MAZ	narrowPeak
ENCSR000EFX	Human	K562	MAZ	narrowPeak
ENCSR000ESL	Mouse	MEL	MAZ	narrowPeak
ENCSR806JZK	Mouse	CH12	MEF2A	narrowPeak
ENCSR000BKB	Human	GM12878	MEF2A	narrowPeak
ENCSR000BNV	Human	K562	MEF2A	narrowPeak

Accession	Species	Cell	TF	Format
ENCSR867SDZ	Mouse	MEL	MEF2A	narrowPeak
ENCSR000ERE	Mouse	CH12	MXI1	narrowPeak
ENCSR000DZI	Human	GM12878	MXI1	narrowPeak
ENCSR000EGZ	Human	K562	MXI1	narrowPeak
ENCSR000ETN	Mouse	MEL	MXI1	narrowPeak
ENCSR000ERN	Mouse	CH12	MYC	narrowPeak
ENCSR000DKU	Human	GM12878	MYC	narrowPeak
ENCSR000FAZ	Human	K562	MYC	narrowPeak
ENCSR000EUA	Mouse	MEL	MYC	narrowPeak
ENCSR980YXJ	Mouse	CH12	NRF1	narrowPeak
ENCSR000DZO	Human	GM12878	NRF1	narrowPeak
ENCSR000EHH	Human	K562	NRF1	narrowPeak
ENCSR135SWH	Mouse	MEL	NRF1	narrowPeak
ENCSR000ERQ	Mouse	CH12	POLR2A	narrowPeak
ENCSR000BGD	Human	GM12878	POLR2A	narrowPeak
ENCSR000BMR	Human	K562	POLR2A	narrowPeak
ENCSR000EUC	Mouse	MEL	POLR2A	narrowPeak
ENCSR000ERH	Mouse	CH12	POLR2AphosphoS2	narrowPeak
ENCSR000DZK	Human	GM12878	POLR2AphosphoS2	narrowPeak
ENCSR000EGF	Human	K562	POLR2AphosphoS2	narrowPeak
ENCSR000ETM	Mouse	MEL	POLR2AphosphoS2	narrowPeak
ENCSR000ERK	Mouse	CH12	RAD21	narrowPeak
ENCSR000BMY	Human	GM12878	RAD21	narrowPeak
ENCSR000BKV	Human	K562	RAD21	narrowPeak
ENCSR000ETS	Mouse	MEL	RAD21	narrowPeak
ENCSR000EQZ	Mouse	CH12	RCOR1	narrowPeak
ENCSR000DZC	Human	GM12878	RCOR1	narrowPeak
ENCSR000EGG	Human	K562	RCOR1	narrowPeak
ENCSR000ESI	Mouse	MEL	RCOR1	narrowPeak
ENCSR000EQY	Mouse	CH12	SIN3A	narrowPeak
ENCSR000DYX	Human	GM12878	SIN3A	narrowPeak
ENCSR000BLR	Human	K562	SIN3A	narrowPeak
ENCSR000ETC	Mouse	MEL	SIN3A	narrowPeak
ENCSR000ERG	Mouse	CH12	SMC3	narrowPeak
ENCSR000DZP	Human	GM12878	SMC3	narrowPeak
ENCSR000EGW	Human	K562	SMC3	narrowPeak
ENCSR000ETL	Mouse	MEL	SMC3	narrowPeak
ENCSR000ERP	Mouse	CH12	TBP	narrowPeak
ENCSR000DZZ	Human	GM12878	TBP	narrowPeak
ENCSR000EHA	Human	K562	TBP	narrowPeak
ENCSR000EUB	Mouse	MEL	TBP	narrowPeak
ENCSR973SOG	Mouse	CH12	USF1	narrowPeak
ENCSR000BGI	Human	GM12878	USF1	narrowPeak
ENCSR000BKT	Human	K562	USF1	narrowPeak
ENCSR705HGT	Mouse	MEL	USF1	narrowPeak
ENCSR000ERJ	Mouse	CH12	USF2	narrowPeak
ENCSR000DZU	Human	GM12878	USF2	narrowPeak
ENCSR000EHG	Human	K562	USF2	narrowPeak
ENCSR000ETF	Mouse	MEL	USF2	narrowPeak
ENCSR000EQO	Mouse	CH12	ZNF384	narrowPeak
ENCSR000DYP	Human	GM12878	ZNF384	narrowPeak
ENCSR000EFP	Human	K562	ZNF384	narrowPeak
ENCSR000ESD	Mouse	MEL	ZNF384	narrowPeak

Table S2

Accession	Species	Cell	format
ENCFF012OHJ	hg19	K562	bigBed
ENCFF836FGC	mm9	MEL	bigBed
ENCFF930VWI	hg19	GM12878	bigBed
ENCFF473FWY	mm9	CH12	bigBed

Table S3

Accession	Species	Cell	Type	Format
ENCFF001YNS	mm9	CH12	Hotspot	bed
ENCFF001YNT	mm9	CH12	Hotspot	bed
ENCFF001WBD	hg19	K562	Hotspot	bed
ENCFF001WNM	hg19	K562	Hotspot	bed
ENCFF001WNL	hg19	K562	Hotspot	bed
ENCFF001WFR	hg19	GM12878	Hotspot	bed
ENCFF001WFS	hg19	GM12878	Hotspot	bed
ENCFF001YRV	mm9	MEL	Hotspot	bed
ENCFF001YRW	mm9	MEL	Hotspot	bed
ENCFF001YLI	mm9	MEL	Hotspot	bed
ENCFF001YRU	mm9	MEL	Hotspot	bed

Table S4

Accession	Species	Cell	format	Lab	Sequencer
ENCFF001MFQ	mm9	CH12	fastq	Hardison	Illumina Genome Analyzer IIx
ENCFF001MFP	mm9	CH12	fastq	Hardison	Illumina Genome Analyzer IIx
ENCFF001MKT	mm9	MEL	fastq	Hardison	HiSeq 2000
ENCFF001MKV	mm9	MEL	fastq	Hardison	HiSeq 2001
ENCFF000DWH	hg19	K562	fastq	Wold	Illumina Genome Analyzer (GAI or GAIIx)
ENCFF000DVT	hg19	K562	fastq	Wold	Illumina Genome Analyzer (GAI or GAIIx)
ENCFF000CXI	hg19	GM12878	fastq	Wold	Illumina Genome Analyzer II
ENCFF000CXH	hg19	GM12878	fastq	Wold	Illumina Genome Analyzer II

Table S5

	Aggregate Grammatical Class	# CRMs	Orthology			Conservation, All Modules			Conservation, Orthologous Modules			Conservation, Unmapped Modules			Conservation, Gains			Conservation, Losses		
			% Lifted	Odds Ratio	Corrected P-Value	% Int Phast	Odds Ratio	Corrected P-Value	% Int Phast	Odds Ratio	Corrected P-Value	% Int Phast	Odds Ratio	Corrected P-Value	% Int Phast	Odds Ratio	Corrected P-Value	% Int Phast	Odds Ratio	Corrected P-Value
Mouse	Lineage-Specific	3219	79.3%	1.95	2.2E-57	56.3%	2.29	2.0E-116	68.0%	1.80	3.3E-44	11.4%	1.55	<3.0E-300	7.8%	1.37	0.06	32.3%	1.28	0.5
	Mouse-Specific	13385	81.9%	2.30	<3.0E-300	66.2%	3.48	<3.0E-300	73.9%	2.40	<3.0E-300	31.6%	5.54	3.2E-228	27.8%	6.29	3.1E-195	51.8%	2.88	8.1E-21
	Human-Specific	138	78.3%	1.83	2.8E-03	63.0%	3.03	1.8E-10	72.2%	2.20	2.9E-04	30.0%	5.14	6.1E-04	30.8%	7.25	<3.0E-300	25.0%	0.89	1.0
	Two-Cell Mixed	2445	88.1%	3.78	1.6E-133	84.2%	9.44	<3.0E-300	88.6%	6.58	5.4E-254	51.4%	12.68	1.7E-83	47.8%	14.94	1.3E-69	64.5%	4.87	5.7E-09
	Three-Cell	13876	86.8%	3.34	<3.0E-300	81.9%	8.01	<3.0E-300	87.4%	5.90	<3.0E-300	45.2%	9.91	<3.0E-300	41.5%	11.57	4.94e-323	63.4%	4.63	6.7E-36
	MCKG	91992	70.7%	1.23	3.6E-96	53.1%	2.01	<3.0E-300	66.1%	1.65	<3.0E-300	21.7%	3.32	<3.0E-300	19.9%	4.06	<3.0E-300	37.4%	1.60	2.9E-16
	Total	**125055**	**74.3%**	**1.47**	**<3.0E-300**	**58.4%**	**2.49**	**<3.0E-300**	**70.3%**	**2.01**	**<3.0E-300**	**23.8%**	**3.75**	**<3.0E-300**	**21.6%**	**4.51**	**<3.0E-300**	**41.5%**	**1.90**	**6.8E-35**
	Background	100000	64.9%	--	--	36.0%	--	--	54.1%	--	--	7.7%	--	--	5.8%	--	--	27.2%	--	--
Human	Lineage-Specific	3356	76.2%	2.18	2.3E-87	66.0%	3.45	2.3E-258	77.6%	1.47	1.0E-15	28.9%	4.17	3.4E-56	12.3%	3.53	2.8E-09	39.4%	4.00	7.1E-41
	Mouse-Specific	176	72.7%	1.81	7.5E-04	68.8%	3.91	3.8E-18	78.1%	1.52	5.3E-02	43.8%	7.95	5.8E-10	36.0%	14.09	6.8E-07	52.2%	6.72	5.1E-05
	Human-Specific	8565	83.2%	3.35	<3.0E-300	84.7%	9.86	<3.0E-300	91.4%	4.52	<3.0E-300	51.7%	10.93	<3.0E-300	40.2%	16.82	3.9E-144	58.7%	8.77	4.6E-192
	Two-Cell Mixed	6287	80.1%	2.72	1.1E-242	80.9%	7.53	<3.0E-300	88.7%	3.34	4.6E-196	50.1%	10.26	1.4E-285	43.0%	18.93	1.0E-156	55.3%	7.61	3.4E-137
	Three-Cell	12687	82.5%	3.21	<3.0E-300	83.8%	9.22	<3.0E-300	90.9%	4.23	<3.0E-300	50.6%	10.47	<3.0E-300	42.4%	18.47	3.9E-234	55.8%	7.76	7.6E-247
	MCKG	53642	74.8%	2.02	<3.0E-300	70.5%	4.26	<3.0E-300	81.2%	1.84	1.0E-293	38.8%	6.49	<3.0E-300	23.6%	7.75	<3.0E-300	47.7%	5.61	<3.0E-300
	Total	**84713**	**77.3%**	**2.31**	**<3.0E-300**	**74.5%**	**5.21**	**<3.0E-300**	**84.3%**	**2.28**	**<3.0E-300**	**41.5%**	**7.25**	**<3.0E-300**	**28.1%**	**9.79**	**<3.0E-300**	**49.5%**	**6.05**	**<3.0E-300**
	Background	69415	58.8%	--	--	36.0%	--	--	70.2%	--	--	8.9%	--	--	3.8%	--	--	14.0%	--	--
All	All	209768	75.5%	1.77	<3.0E-300	64.9%	3.29	<3.0E-300	76.1%	2.10	<3.0E-300	30.4%	4.87	<3.0E-300	22.9%	3.52	<3.0E-300	47.7%	4.62	<3.0E-300
	Background	169415	62.4%	--	--	36.0%	--	--	60.3%	--	--	8.2%	--	--	5.2%	--	--	16.5%	--	--

Table S6

	Aggregate Grammatical Class	# CRMs	DHS All				DHS Orth				DHS Nonorth				DHS Gain				DHS Loss			
			% DHS	% DHS BG	Odds Ratio	Corrected P-Value	%DHS	% DHS BG	Odds Ratio	Corrected P-Value	%DHS	% DHS BG	Odds Ratio	Corrected P-Value	%DHS	% DHS BG	Odds Ratio	Corrected P-Value	%DHS	% DHS BG	Odds Ratio	Corrected P-Value
Mouse	Lineage-Specific	3219	95.78%	29.64%	53.74	< 1.9E-283	96.56%	4.70%	56.68	< 1.9E-283	92.19%	20.00%	46.88	3.9E-96	91.18%	20.50%	39.71	1.3E-78	97.98%	16.67%	213.28	5.8E-19
	Mouse-Specific	13385	98.31%	26.40%	162.26	< 1.9E-283	98.54%	2.04%	163.54	< 1.9E-283	97.16%	19.59%	139.60	< 1.9E-283	97.02%	19.62%	132.22	< 1.9E-283	97.89%	19.42%	187.60	4.2E-63
	Human-Specific	138	87.68%	25.00%	20.98	2.1E-24	87.00%	16.46%	19.29	1.8E-17	86.67%	42.86%	7.96	2.7E-02	88.46%	33.33%	13.35	1.2E-02	75.00%	100.00%	0.00	1
	Two-Cell Mixed	2445	99.75%	35.87%	724.79	< 1.9E-283	99.80%	0.28%	783.02	< 1.9E-283	99.31%	27.91%	358.76	9.1E-48	99.12%	30.88%	242.94	4.8E-35	100.00%	16.67%	Inf	3.7E-13
	Three-Cell	13876	99.51%	34.41%	384.87	< 1.9E-283	99.59%	0.56%	416.00	< 1.9E-283	98.91%	26.19%	254.12	2.6E-264	99.01%	25.80%	286.39	5.0E-226	98.41%	28.36%	150.62	9.3E-39
	MCKG	91992	80.35%	21.70%	14.76	< 1.9E-283	83.04%	21.96%	15.37	< 1.9E-283	73.58%	15.23%	15.50	< 1.9E-283	72.88%	14.98%	15.24	< 1.9E-283	79.96%	17.49%	18.80	1.9E-239
	Total	**125055**	**85.18%**	**24.04%**	**18.16**	**< 1.9E-283**	**87.72%**	**16.21%**	**19.50**	**< 1.9E-283**	**77.44%**	**16.25%**	**17.69**	**< 1.9E-283**	**76.58%**	**16.00%**	**17.17**	**< 1.9E-283**	**84.37%**	**18.40%**	**23.90**	**< 1.9E-283**
Human	Lineage-Specific	3356	99.37%	25.58%	456.04	< 1.9E-283	99.57%	0.67%	528.86	< 1.9E-283	98.75%	15.53%	421.97	2.4E-191	97.73%	14.05%	255.68	3.7E-90	99.39%	12.97%	1027.97	9.6E-124
	Mouse-Specific	176	98.86%	26.58%	235.10	1.6E-50	99.21%	1.04%	309.40	1.9E-33	97.92%	7.14%	373.16	4.6E-11	100.00%	0.00%	Inf	2.7E-06	95.65%	16.67%	69.96	5.9E-04
	Human-Specific	8565	99.31%	34.82%	271.45	< 1.9E-283	99.49%	0.83%	298.47	< 1.9E-283	98.54%	23.76%	215.67	2.7E-259	97.09%	15.23%	183.26	5.9E-145	99.44%	22.85%	591.34	2.4E-175
	Two-Cell Mixed	6287	99.90%	30.57%	2285.62	< 1.9E-283	99.92%	0.12%	2242.03	< 1.9E-283	99.84%	21.35%	2264.44	5.1E-279	99.62%	10.80%	1997.39	2.8E-173	100.00%	24.39%	Inf	1.9E-158
	Three-Cell	12687	99.68%	34.38%	586.97	< 1.9E-283	99.84%	0.27%	923.32	< 1.9E-283	99.05%	21.27%	384.74	< 1.9E-283	98.02%	13.26%	320.61	3.3E-243	99.71%	20.87%	1228.72	1.9E-283
	MCKG	53642	88.32%	21.19%	28.12	< 1.9E-283	89.66%	15.84%	25.69	< 1.9E-283	84.84%	16.58%	28.13	< 1.9E-283	82.00%	9.52%	43.24	< 1.9E-283	86.49%	16.77%	31.79	< 1.9E-283
	Total	**84713**	**92.45%**	**25.09%**	**36.57**	**< 1.9E-283**	**93.57%**	**9.97%**	**34.19**	**< 1.9E-283**	**89.09%**	**17.73%**	**37.82**	**< 1.9E-283**	**87.07%**	**10.49%**	**57.40**	**< 1.9E-283**	**90.30%**	**17.86%**	**42.77**	**< 1.9E-283**
	Total All	209768	88.12%	24.47%	22.89	< 1.9E-283	90.19%	13.82%	23.66	< 1.9E-283	81.80%	16.89%	22.11	< 1.9E-283	78.70%	14.09%	22.54	< 1.9E-283	88.95%	17.96%	36.74	< 1.9E-283

Table S7

	Aggregate Grammatical Class	CRMs	ACS All				ACS Orth				ACS Nonorth				ACS Gain				ACS Loss			
			% ACS	% ACS BG	Odds Ratio	Corrected P-Value	%ACS	% ACS BG	Odds Ratio	Corrected P-Value	%ACS	% ACS BG	Odds Ratio	Corrected P-Value	%ACS	% ACS BG	Odds Ratio	Corrected P-Value	%ACS	% ACS BG	Odds Ratio	Corrected P-Value
Mouse	Lineage-Specific	3219	77.20%	25.62%	9.82	< 2E-272	80.65%	28.94%	10.23	< 2E-272	61.71%	18.70%	6.99	7.42E-30	59.08%	21.00%	5.42	2.71E-20	76.77%	3.33%	92.18	6.83E-13
	Mouse-Specific	13385	88.03%	27.70%	19.19	< 2E-272	89.65%	31.15%	19.15	< 2E-272	80.05%	19.59%	16.45	2.46E-186	78.74%	19.97%	14.83	3.13E-148	87.11%	17.48%	31.52	2.51E-40
	Human-Specific	138	65.22%	35.19%	3.44	3.35E-06	66.00%	40.51%	2.83	3.35E-06	60.00%	0.00%	Inf	7.99E-03	65.38%	0.00%	Inf	5.52E-03	25.00%	0.00%	Inf	1
	Two-Cell Mixed	2445	94.97%	32.22%	39.65	< 2E-272	95.82%	35.51%	41.56	< 2E-272	88.28%	24.42%	22.98	8.96E-29	87.28%	27.94%	17.43	1.04E-19	91.94%	11.11%	80.34	3.96E-10
	Three-Cell	13876	88.41%	29.80%	17.97	< 2E-272	89.65%	31.94%	18.46	< 2E-272	77.44%	24.83%	10.38	1.04E-93	76.40%	25.80%	9.29	9.84E-73	82.48%	19.40%	19.34	3.49E-22
	MCKG	91992	33.11%	25.18%	1.47	2.35E-272	37.03%	28.35%	1.49	2.35E-272	22.44%	19.91%	1.16	1.51E-06	21.56%	19.84%	1.11	2.56E-03	30.32%	20.57%	1.68	6.76E-08
Human	Lineage-Specific	3356	92.73%	28.56%	31.89	< 2E-272	93.22%	33.05%	27.81	< 2E-272	90.98%	20.19%	39.67	6.94E-120	87.66%	18.98%	29.95	5.06E-45	93.06%	21.08%	49.56	1.55E-75
	Mouse-Specific	176	95.45%	32.28%	43.38	4.31E-37	94.44%	35.42%	30.36	4.31E-37	97.92%	21.43%	138.15	2.44E-08	100.00%	25.00%	Inf	7.58E-05	95.65%	16.67%	69.96	5.85E-04
	Human-Specific	8565	96.12%	36.90%	42.37	< 2E-272	96.55%	41.12%	40.00	< 2E-272	94.31%	27.48%	43.60	3.24E-186	93.82%	25.82%	43.13	8.34E-72	94.62%	28.48%	43.90	5.22E-115
	Two-Cell Mixed	6287	91.55%	33.78%	21.24	< 2E-272	92.48%	37.65%	20.37	< 2E-272	88.04%	25.37%	21.59	1.40E-137	87.78%	27.42%	18.89	5.61E-52	88.23%	24.04%	23.55	4.05E-86
	Three-Cell	12687	90.95%	34.89%	18.76	< 2E-272	91.65%	39.36%	16.91	< 2E-272	87.77%	23.85%	22.88	2.31E-230	88.84%	23.74%	25.44	2.33E-92	87.09%	23.91%	21.41	1.20E-138
	MCKG	53642	37.48%	28.19%	1.53	1.42E-212	39.59%	32.61%	1.35	1.42E-212	31.74%	23.62%	1.50	4.34E-29	28.35%	22.21%	1.39	3.51E-07	33.72%	24.44%	1.57	3.04E-23
All	**Total**	**209768**	**51.63%**	**27.84%**	**2.77**	**< 2E-272**	**55.86%**	**31.46%**	**2.76**	**< 2E-272**	**38.03%**	**21.72%**	**2.21**	**< 2E-272**	**33.46%**	**20.81%**	**1.91**	**2.47E-154**	**48.56%**	**23.57%**	**3.06**	**1.49E-248**
	Total.hg19	**84713**	**44.91%**	**27.69%**	**2.13**	**< 2E-272**	**48.99%**	**31.42%**	**2.10**	**< 2E-272**	**33.58%**	**21.76%**	**1.82**	**8.23E-180**	**27.56%**	**20.64%**	**1.46**	**1.95E-45**	**46.47%**	**23.89%**	**2.77**	**1.68E-196**
	Total.mm9	**125055**	**47.50%**	**26.10%**	**2.56**	**< 2E-272**	**52.47%**	**29.28%**	**2.67**	**< 2E-272**	**31.36%**	**20.11%**	**1.82**	**5.40E-109**	**29.87%**	**20.17%**	**1.69**	**8.48E-74**	**43.40%**	**19.53%**	**3.16**	**3.75E-47**

Table S8

#GWAS Ontology Term	Category
aggressive behavior	behavior
alcohol dependence	behavior
alcohol drinking	behavior
behavior or behavioral disorder measurement	behavior
coffee consumption	behavior
cups of coffee per day measurement	behavior
illegal drug consumption	behavior
nicotine use	behavior
non-substance related disinhibited behaviour	behavior
smoking behavior	behavior
smoking behaviour measurement	behavior
smoking cessation	behavior
smoking initiation	behavior
aging	biological_process
energy expenditure	biological_process
energy intake	biological_process
physical activity	biological_process
sneezing	biological_process
albumin:globulin ratio measurement	blood
alpha globulin measurement	blood
blood metabolite measurement	blood
blood sedimentation	blood
coagulation factor measurement	blood
erythrocyte count	blood
erythrocyte measurement	blood
factor VIII measurement	blood
fetal hemoglobin measurement	blood
fibrinogen measurement	blood
hematocrit	blood
hematological measurement	blood
hemoglobin A2 measurement	blood
hemoglobin E disease	blood
hemoglobin measurement	blood
mean corpuscular hemoglobin	blood
mean corpuscular hemoglobin concentration	blood
mean corpuscular volume	blood
mean platelet volume	blood
partial thromboplastin time	blood
platelet aggregation	blood
platelet count	blood
protein c measurement	blood
red blood cell distribution width	blood
sickle cell anemia	blood
total blood protein measurement	blood
von Willebrand factor measurement	blood
ankle brachial index	body_measurement
anthropometry	body_measurement
arm span	body_measurement
Attached earlobe	body_measurement
axial length measurement	body_measurement
birth weight	body_measurement
body composition measurement	body_measurement
body fat distribution	body_measurement
body height	body_measurement
body mass index	body_measurement
body weight	body_measurement
dense area measurement	body_measurement
foot	body_measurement
head circumference	body_measurement
height growth measurement	body_measurement
height-adjusted body mass index	body_measurement
hip bone size	body_measurement
hip circumference	body_measurement
hip geometry	body_measurement
infant body height	body_measurement
infant head circumference	body_measurement
intracranial volume	body_measurement
lean body mass	body_measurement
mammographic density measurement	body_measurement
mammographic density percentage	body_measurement
metabolic rate measurement	body_measurement
non-dense area measurement	body_measurement
spine bone size	body_measurement
subcutaneous adipose tissue measurement	body_measurement
visceral adipose tissue measurement	body_measurement

#GWAS Ontology Term	Category
waist circumference	body_measurement
waist-hip ratio	body_measurement
bone density	bone
braces	bone
osteitis deformans	bone
scoliosis	bone
alpha peak frequency measurement	brain
alpha wave measurement	brain
beta wave measurement	brain
brain measurement	brain
brain serotonin transporter measurement	brain
cognition	brain
cortical thickness	brain
delta wave measurement	brain
electroencephalogram measurement	brain
frontal theta oscillation measurement	brain
Heschl's gyrus morphology measurement	brain
hippocampal volume	brain
information processing speed	brain
intelligence	brain
lentiform nucleus volume	brain
mathematical ability	brain
memory performance	brain
memory, short-term	brain
mental process	brain
neurofibrilliary tangles measurement	brain
paragraph delayed recall measurement	brain
reasoning	brain
self reported educational attainment	brain
superior frontal gyrus grey matter volume measu	brain
theta wave measurement	brain
volumetric brain mri	brain
white matter integrity	brain
white matter microstructure measurement	brain
word list delayed recall measurement	brain
word reading	brain
Alzheimer's disease neuropathologic change	brain_disease
Alzheimers disease	brain_disease
anorexia nervosa	brain_disease
anxiety	brain_disease
anxiety disorder	brain_disease
asperger syndrome	brain_disease
attention deficit hyperactivity disorder	brain_disease
autism	brain_disease
autism spectrum disorder	brain_disease
autism spectrum disorder symptom	brain_disease
bipolar disorder	brain_disease
bulimia nervosa	brain_disease
conduct disorder	brain_disease
dementia	brain_disease
drug dependence	brain_disease
dyslexia	brain_disease
eating disorder	brain_disease
epilepsy	brain_disease
hippocampal atrophy	brain_disease
HVA measurement	brain_disease
insomnia	brain_disease
language impairment	brain_disease
Lewy body dementia	brain_disease
Lewy body dementia measurement	brain_disease
mental or behavioural disorder	brain_disease
MHPG measurement	brain_disease
migraine disorder	brain_disease
migraine with aura	brain_disease
migraine without aura	brain_disease
mood disorder	brain_disease
narcolepsy	brain_disease
neuritic plaque measurement	brain_disease
neuropsychological test	brain_disease
nicotine dependence	brain_disease
p-tau measurement	brain_disease
panic disorder	brain_disease
parkinson's disease	brain_disease
plasma beta-amyloid 1-40 measurement	brain_disease
plasma beta-amyloid 1-42 measurement	brain_disease
post-traumatic stress disorder	brain_disease

#GWAS Ontology Term	Category
progressive supranuclear palsy	brain_disease
psychosis	brain_disease
psychotic symptoms	brain_disease
schizoaffective disorder	brain_disease
schizophrenia	brain_disease
social communication impairment	brain_disease
suicidal ideation	brain_disease
unipolar depression	brain_disease
acute lymphoblastic leukemia	cancer
B-cell acute lymphoblastic leukemia	cancer
bladder carcinoma	cancer
breast carcinoma	cancer
cancer biomarker measurement	cancer
central nervous system cancer	cancer
cervical carcinoma	cancer
chronic lymphocytic leukemia	cancer
colorectal cancer	cancer
diffuse large b-cell lymphoma	cancer
Esophageal adenocarcinoma	cancer
esophageal carcinoma	cancer
esophageal squamous cell carcinoma	cancer
gastric carcinoma	cancer
hepatocellular carcinoma	cancer
hodgkins lymphoma	cancer
igfbp-3 measurement	cancer
lung adenocarcinoma	cancer
lung carcinoma	cancer
lymphoma	cancer
malignant epithelial tumor of ovary	cancer
marginal zone B-cell lymphoma	cancer
melanoma	cancer
multiple myeloma	cancer
neoplasm of mature b-cells	cancer
nephroblastoma	cancer
neuroblastoma	cancer
nevus	cancer
nodular sclerosis Hodgkin lymphoma	cancer
non-small cell lung carcinoma	cancer
osteosarcoma	cancer
ovarian carcinoma	cancer
pancreatic carcinoma	cancer
prostate carcinoma	cancer
prostate specific antigen measurement	cancer
renal cell carcinoma	cancer
serum carcinoembryonic antigen measurement	cancer
squamous cell carcinoma	cancer
testicular carcinoma	cancer
triple-negative breast cancer	cancer
abdominal aortic artery calcification	cardiovascular
aortic root size	cardiovascular
asymmetrical dimethylarginine measurement	cardiovascular
atrial fibrillation	cardiovascular
blood pressure	cardiovascular
brain aneurysm	cardiovascular
Brugada syndrome	cardiovascular
cardiac hypertrophy	cardiovascular
cardiac troponin T measurement	cardiovascular
cardiovascular disease	cardiovascular
cardiovascular measurement	cardiovascular
carotid artery disease	cardiovascular
cis/trans-18:2 fatty acid measurement	cardiovascular
common carotid intimal medial thickness	cardiovascular
conotruncal heart defect	cardiovascular
coronary artery calcification	cardiovascular
coronary heart disease	cardiovascular
creatine kinase measurement	cardiovascular
diastolic blood pressure	cardiovascular
dihydroxy docosatrienoic acid measurement	cardiovascular
ejection fraction measurement	cardiovascular
electrocardiography	cardiovascular
exercise test	cardiovascular
heart failure	cardiovascular
heart function measurement	cardiovascular
heart rate	cardiovascular
Hydroxy-leucine measurement	cardiovascular
hypertension	cardiovascular

#GWAS Ontology Term	Category
icam-1 measurement	cardiovascular
internal carotid intimal medial thickness	cardiovascular
large artery stroke	cardiovascular
lipoprotein-associated phospholipase a(2) measur	cardiovascular
myeloperoxidase measurement	cardiovascular
myocardial infarction	cardiovascular
nt-probnp measurement	cardiovascular
P wave duration	cardiovascular
pr interval	cardiovascular
PR segment	cardiovascular
pulmonary artery enlargement	cardiovascular
pulmonary function measurement	cardiovascular
pyroglutamine measurement	cardiovascular
QRS complex	cardiovascular
QRS duration	cardiovascular
qt interval	cardiovascular
resting heart rate	cardiovascular
rr interval	cardiovascular
serum homoarginine measurement	cardiovascular
serum ST2 measurement	cardiovascular
soluble p-selectin measurement	cardiovascular
stroke	cardiovascular
sudden cardiac arrest	cardiovascular
systolic blood pressure	cardiovascular
tissue plasminogen activator measurement	cardiovascular
total trans-18:1 fatty acid measurement	cardiovascular
trans fatty acid measurement	cardiovascular
trans-16:1n-7 fatty acid measurement	cardiovascular
trans/cis-18:2 fatty acid measurement	cardiovascular
trans/trans-18:2 fatty acid measurement	cardiovascular
triglyceride measurement	cardiovascular
uric acid measurement	cardiovascular
venous thromboembolism	cardiovascular
androgenetic alopecia	common_traits
eye color	common_traits
hair color	common_traits
hair morphology	common_traits
personality	common_traits
personality trait	common_traits
temperament and character inventory	common_traits
dental caries	dental
molar-incisor hypomineralization	dental
odontogenesis	dental
periodontitis	dental
pit and fissure surface dental caries	dental
smooth surface dental caries	dental
tooth agenesis	dental
tooth eruption	dental
neonate	developmental_stage
a1c measurement	diabetes
acute insulin response measurement	diabetes
disposition index measurement	diabetes
fasting blood glucose measurement	diabetes
fasting blood insulin measurement	diabetes
glucose effectiveness measurement	diabetes
glucose homeostasis measurement	diabetes
HOMA-IR	diabetes
insulin metabolic clearance rate measurement	diabetes
insulin resistance	diabetes
insulin sensitivity measurement	diabetes
resistin measurement	diabetes
type 1 diabetes nephropathy	diabetes
type ii diabetes mellitus	diabetes
aspirin induced asthma	drug_response
chemotherapy-induced hypertension	drug_response
cumulative dose response to bevacizumab	drug_response
drug-induced agranulocytosis	drug_response
drug-induced liver injury	drug_response
response to anticoagulant	drug_response
response to anticonvulsant	drug_response
response to antidepressant	drug_response
response to antimicrotubule agent	drug_response
response to antineoplastic agent	drug_response
response to antipsychotic drug	drug_response
response to bevacizumab	drug_response
response to bronchodilator	drug_response

#GWAS Ontology Term	Category
response to candesartan	drug_response
response to carboplatin	drug_response
response to clozapine	drug_response
response to corticosteroid	drug_response
response to cyclophosphamide	drug_response
response to cytosine arabinoside	drug_response
response to docetaxel	drug_response
response to drug	drug_response
response to efavirenz	drug_response
response to etoposide	drug_response
response to fenofibrate	drug_response
response to gemcitabine	drug_response
response to haloperidol	drug_response
response to hydrochlorothiazide	drug_response
response to interferon beta	drug_response
response to irinotecan	drug_response
response to methotrexate	drug_response
response to methylphenidate	drug_response
response to paclitaxel	drug_response
response to platinum based chemotherapy	drug_response
response to reverse transcriptase inhibitor	drug_response
response to selective serotonin reuptake inhibitor	drug_response
response to statin	drug_response
response to vitamin	drug_response
Abnormality of refraction	eye
age-related cataract	eye
age-related macular degeneration	eye
Astigmatism	eye
contrast sensitivity measurement	eye
corneal topography	eye
diabetic retinopathy	eye
exfoliation syndrome	eye
eye measurement	eye
hypermetropia	eye
intraocular pressure measurement	eye
Myopia	eye
open-angle glaucoma	eye
optic disc size measurement	eye
pathological myopia	eye
retinopathy	eye
rhegmatogenous retinal detachment	eye
Barrett's esophagus	gut
celiac disease	gut
Hirschsprung disease	gut
sclerosing cholangitis	gut
cortisol secretion measurement	hormone_measurement
estradiol measurement	hormone_measurement
hormone measurement	hormone_measurement
insulin measurement	hormone_measurement
leptin measurement	hormone_measurement
leptin receptor measurement	hormone_measurement
parathyroid hormone measurement	hormone_measurement
sex hormone globulin binding measurement	hormone_measurement
thyroid stimulating hormone measurement	hormone_measurement
thyroxine measurement	hormone_measurement
acne	immune_system
aids	immune_system
allergic sensitization measurement	immune_system
allergy	immune_system
alloimmunization	immune_system
alopecia areata	immune_system
ankylosing spondylitis	immune_system
anti-Heliobacter pylori serum IgG measurement	immune_system
anti-neutrophil antibody associated vasculitis	immune_system
antibody measurement	immune_system
antiphospholipid antibody measurement	immune_system
atopic eczema	immune_system
atopy	immune_system
autoantibody measurement	immune_system
autoimmune hepatits type 1	immune_system
Behcet's disease	immune_system
ccl2 measurement	immune_system
CCL4 measurement	immune_system
CCL5 measurement	immune_system
chronic childhood arthritis	immune_system
chronic hepatitis B infection	immune_system

#GWAS Ontology Term	Category
chronic hepatitis C infection	immune_system
complement C3 measurement	immune_system
complement C4 measurement	immune_system
crohn's disease	immune_system
eosinophil count	immune_system
eosinophilic esophagitis	immune_system
Epstein-Barr virus infection	immune_system
graves disease	immune_system
hepatitis B infection	immune_system
HIV viral set point measurement	immune_system
HIV-1 infection	immune_system
iga glomerulonephritis	immune_system
IgF-1 measurement	immune_system
immune system disease	immune_system
inflammatory bowel disease	immune_system
interleukin 1 receptor antagonist measurement	immune_system
interleukin 12 measurement	immune_system
interleukin 18 measurement	immune_system
interleukin-1 beta measurement	immune_system
interleukin-6 measurement	immune_system
interleukin-6 receptor measurement	immune_system
interleukin-8 measurement	immune_system
irritable bowel syndrome	immune_system
leukocyte count	immune_system
malaria	immune_system
monocyte count	immune_system
monocyte early outgrowth colony forming unit	immune_system
multiple sclerosis	immune_system
neutrophil count	immune_system
osteoarthritis	immune_system
parasitemia measurement	immune_system
psoriasis	immune_system
pyoderma gangrenosum	immune_system
response to dietary antigen	immune_system
response to vaccine	immune_system
response to virus	immune_system
rheumatoid arthritis	immune_system
seasonal allergic rhinitis	immune_system
serum IgE measurement	immune_system
serum IgG glycosylation measurement	immune_system
serum IgG measurement	immune_system
serum IgM measurement	immune_system
Sjogren syndrome	immune_system
Staphylococcus aureus infection	immune_system
systemic lupus erythematosus	immune_system
systemic scleroderma	immune_system
thyroid peroxidase antibody measurement	immune_system
Trypanosoma cruzi seropositivity	immune_system
tuberculosis	immune_system
type i diabetes mellitus	immune_system
ulcerative colitis	immune_system
virologic response measurement	immune_system
vitiligo	immune_system
apolipoprotein a 1 measurement	inflammatory_measurement
c-reactive protein measurement	inflammatory_measurement
cytokine measurement	inflammatory_measurement
interleukin 10 measurement	inflammatory_measurement
l lactate dehydrogenase measurement	inflammatory_measurement
serum amyloid a protein measurement	inflammatory_measurement
transforming growth factor beta measurement	inflammatory_measurement
tumor necrosis factor-alpha measurement	inflammatory_measurement
ykl40 measurement	inflammatory_measurement
alpha macroglobulin measurement	kidney
blood urea nitrogen measurement	kidney
chronic kidney disease	kidney
diabetic nephropathy	kidney
kidney disease	kidney
membranous glomerulonephritis	kidney
nephrotic syndrome	kidney
renal system measurement	kidney
serum creatinine measurement	kidney
alkaline phosphatase measurement	liver
aspartate aminotransferase measurement	liver
biliary atresia	liver
biliary liver cirrhosis	liver
bilirubin measurement	liver

#GWAS Ontology Term	Category
butyrylcholinesterase measurement	liver
hepatitis C induced liver cirrhosis	liver
liver enzyme measurement	liver
non-alcoholic fatty liver disease	liver
serum alanine aminotransferase measurement	liver
serum albumin measurement	liver
serum gamma-glutamyl transferase measurement	liver
airway responsiveness measurement	lung
airway wall thickness measurement	lung
asthma	lung
bronchopulmonary dysplasia	lung
CC16 measurement	lung
childhood onset asthma	lung
chronic bronchitis	lung
chronic obstructive pulmonary disease	lung
cystic fibrosis	lung
emphysema	lung
fev/fec ratio	lung
forced expiratory volume	lung
idiopathic pulmonary fibrosis	lung
interstitial lung disease	lung
maximal midexpiratory flow rate	lung
maximal oxygen uptake measurement	lung
surfactant protein D measurement	lung
vital capacity	lung
5-HIAA measurement	nervous_system
age-related hearing impairment	nervous_system
amyotrophic lateral sclerosis	nervous_system
circadian rhythm	nervous_system
febrile seizures	nervous_system
MMR-related febrile seizures	nervous_system
sensory perception of smell	nervous_system
sensory perception of taste	nervous_system
sporadic amyotrophic lateral sclerosis	nervous_system
cleft lip	other_disease
dupuytren contracture	other_disease
leprosy	other_disease
lumbar disc degeneration	other_disease
mucocutaneous lymph node syndrome	other_disease
orofacial clefting syndrome	other_disease
pancreatitis	other_disease
prion disease	other_disease
sarcoidosis	other_disease
stevens-johnson syndrome	other_disease
adhesion molecule measurement	other_measurement
adiponectin measurement	other_measurement
age at onset	other_measurement
amino acid measurement	other_measurement
antioxidant measurement	other_measurement
antisaccade response measurement	other_measurement
arachidonic acid measurement	other_measurement
asbestos exposure measurement	other_measurement
beta-2 microglobulin measurement	other_measurement
calcium measurement	other_measurement
cotinine glucuronidation measurement	other_measurement
DNA methylation	other_measurement
docosapentaenoic acid measurement	other_measurement
economic and social preference	other_measurement
electrodermal activity measurement	other_measurement
event free survival time	other_measurement
fatty acid measurement	other_measurement
ferritin measurement	other_measurement
folic acid measurement	other_measurement
fructose-bisphosphate aldolase measurement	other_measurement
functional laterality	other_measurement
genetic variation	other_measurement
glycoprotein measurement	other_measurement
high density lipoprotein cholesterol measurement	other_measurement
homocysteine measurement	other_measurement
IGFBP-1 measurement	other_measurement
insulin like growth factor measurement	other_measurement
iron biomarker measurement	other_measurement
linoleic acid measurement	other_measurement
linolenic acid measurement	other_measurement
lipid measurement	other_measurement
longevity	other_measurement

#GWAS Ontology Term	Category
low density lipoprotein cholesterol measurement	other_measurement
magnesium measurement	other_measurement
matrix metalloproteinase measurement	other_measurement
metabolite measurement	other_measurement
mitochondrial DNA measurement	other_measurement
mortality	other_measurement
myoglobin measurement	other_measurement
N-glycan measurement	other_measurement
nitric oxide exhalation measurement	other_measurement
oligoclonal band measurement	other_measurement
omega-6 polyunsaturated fatty acid measurement	other_measurement
phospholipid measurement	other_measurement
phosphorus measurement	other_measurement
plasminogen activator inhibitor 1 measurement	other_measurement
protein measurement	other_measurement
selenium measurement	other_measurement
self rated health	other_measurement
serum copper measurement	other_measurement
serum dimethylarginine measurement	other_measurement
serum iron measurement	other_measurement
serum metabolite measurement	other_measurement
serum selenium measurement	other_measurement
serum zinc measurement	other_measurement
short sleep	other_measurement
skin conductance level	other_measurement
skin conductance response amplitude	other_measurement
skin conductance response frequency	other_measurement
sleep duration	other_measurement
sleep measurement	other_measurement
soluble transferrin receptor measurement	other_measurement
sphingolipid measurement	other_measurement
survival time	other_measurement
telomere length	other_measurement
total cholesterol measurement	other_measurement
transferrin measurement	other_measurement
transferrin saturation measurement	other_measurement
urate measurement	other_measurement
urinary metabolite measurement	other_measurement
urinary uromodulin measurement	other_measurement
very long-chain saturated fatty acid measurement	other_measurement
vitamin B12 measurement	other_measurement
vitamin D measurement	other_measurement
vitamin measurement	other_measurement
metabolic syndrome	other_metabolic_disorder
obesity	other_metabolic_disorder
overweight body mass index status	other_metabolic_disorder
age at menarche	reproductive
age at menopause	reproductive
azoospermia	reproductive
endometriosis	reproductive
erectile dysfunction	reproductive
follicle stimulating hormone measurement	reproductive
gestational age	reproductive
hypospadias	reproductive
male infertility	reproductive
polycystic ovary syndrome	reproductive
preeclampsia	reproductive
puberty	reproductive
puberty onset measurement	reproductive
sexual dysfunction	reproductive
testosterone measurement	reproductive
uterine fibroid	reproductive
suntan	skin
toxic epidermal necrolysis	skin
response to cold pressor test	stimulus_response
response to radiation	stimulus_response
response to red blood cell transfusion	stimulus_response

Table S9

GWAS Ontology Term	Observed Peak Count	Expected Peak Count	Expected Frequency	Observed Frequency	Freq FC	Raw P-Value	Corrected. P-Value (fdr)	Category
inflammatory bowel disease	48	15	0.007628	0.023821	3.12	1.74E-11	8.44E-09	immune_system
immune system disease	16	2	0.001165	0.007940	6.82	4.29E-09	1.04E-06	immune_system
erythrocyte measurement	23	5	0.002620	0.011414	4.36	9.75E-09	1.58E-06	blood
multiple sclerosis	56	25	0.012170	0.027792	2.28	2.80E-08	2.98E-06	immune_system
systemic lupus erythematosus	42	16	0.007861	0.020844	2.65	3.06E-08	2.98E-06	immune_system
crohn's disease	54	25	0.012287	0.026799	2.18	2.03E-07	1.64E-05	immune_system
systemic scleroderma	18	4	0.002213	0.008933	4.04	1.09E-06	7.59E-05	immune_system
anti-Heliobacter pylori serum IgG measurement	5	0	0.000116	0.002481	21.31	4.86E-06	2.62E-04	immune_system
self reported educational attainment	24	8	0.004076	0.011911	2.92	5.40E-06	2.62E-04	brain
ulcerative colitis	36	15	0.007628	0.017866	2.34	4.54E-06	2.62E-04	immune_system
mood disorder	4	0	0.000058	0.001985	34.09	7.17E-06	3.17E-04	brain_disease
esophageal squamous cell carcinoma	8	1	0.000582	0.003970	6.82	3.13E-05	1.27E-03	cancer
rheumatoid arthritis	53	29	0.014441	0.026303	1.82	3.81E-05	1.42E-03	immune_system
seasonal allergic rhinitis	15	4	0.002155	0.007444	3.46	4.84E-05	1.68E-03	immune_system
parkinson's disease	29	13	0.006347	0.014392	2.27	6.37E-05	2.06E-03	brain_disease
metabolite measurement	24	10	0.004833	0.011911	2.46	7.74E-05	2.35E-03	other_measurement
blood metabolite measurement	47	26	0.012694	0.023325	1.84	8.30E-05	2.37E-03	blood
intracranial volume	4	0	0.000116	0.001985	17.05	1.05E-04	2.82E-03	body_measurement
schizophrenia	67	41	0.020439	0.033251	1.63	1.14E-04	2.91E-03	brain_disease
platelet count	25	11	0.005357	0.012407	2.32	1.43E-04	3.48E-03	blood
colorectal cancer	23	10	0.004950	0.011414	2.31	2.74E-04	6.33E-03	cancer
progressive supranuclear palsy	7	1	0.000641	0.003474	5.42	3.84E-04	8.49E-03	brain_disease
psoriasis	14	5	0.002387	0.006948	2.91	4.74E-04	0.010	immune_system
platelet aggregation	12	4	0.001863	0.005955	3.20	5.26E-04	0.011	blood
functional laterality	8	2	0.000932	0.003970	4.26	7.29E-04	0.014	other_measurement
allergy	12	4	0.002038	0.005955	2.92	1.12E-03	0.021	immune_system
urate measurement	18	8	0.003901	0.008933	2.29	1.29E-03	0.023	other_measurement
cardiac troponin T measurement	8	2	0.001048	0.003970	3.79	1.53E-03	0.026	cardiovascular
total blood protein measurement	5	1	0.000408	0.002481	6.09	1.58E-03	0.026	blood
coronary heart disease	30	17	0.008327	0.014888	1.79	2.17E-03	0.035	cardiovascular
adhesion molecule measurement	4	1	0.000291	0.001985	6.82	3.10E-03	0.046	other_measurement
nevus	4	1	0.000291	0.001985	6.82	3.10E-03	0.046	cancer
soluble p-selectin measurement	4	1	0.000291	0.001985	6.82	3.10E-03	0.046	cardiovascular
magnesium measurement	5	1	0.000524	0.002481	4.73	4.58E-03	0.064	other_measurement
response to irinotecan	5	1	0.000524	0.002481	4.73	4.58E-03	0.064	drug_response
mean corpuscular volume	17	8	0.004134	0.008437	2.04	5.38E-03	0.073	blood
Brugada syndrome	3	0	0.000175	0.001489	8.52	5.59E-03	0.073	cardiovascular
vitiligo	11	4	0.002213	0.005459	2.47	6.19E-03	0.079	immune_system
serum homoarginine measurement	2	0	0.000058	0.000993	17.05	6.37E-03	0.079	cardiovascular
aggressive behavior	6	2	0.000815	0.002978	3.65	6.80E-03	0.081	behavior
osteoarthritis	6	2	0.000815	0.002978	3.65	6.80E-03	0.081	immune_system
body height	203	172	0.085250	0.100744	1.18	8.20E-03	0.095	body_measurement
chronic lymphocytic leukemia	11	5	0.002329	0.005459	2.34	8.86E-03	0.100	cancer
protein c measurement	4	1	0.000408	0.001985	4.87	9.90E-03	0.109	blood
testicular carcinoma	8	3	0.001456	0.003970	2.73	1.05E-02	0.113	cancer
mean platelet volume	13	6	0.003086	0.006452	2.09	1.15E-02	0.121	blood
rhegmatogenous retinal detachment	3	0	0.000233	0.001489	6.39	1.22E-02	0.126	eye
myeloperoxidase measurement	4	1	0.000466	0.001985	4.26	1.54E-02	0.156	cardiovascular
cups of coffee per day measurement	4	1	0.000524	0.001985	3.79	2.26E-02	0.219	behavior
vitamin measurement	4	1	0.000524	0.001985	3.79	2.26E-02	0.219	other_measurement
asbestos exposure measurement	2	0	0.000116	0.000993	8.52	2.36E-02	0.225	other_measurement
personality trait	10	5	0.002387	0.004963	2.08	2.53E-02	0.237	common_traits
complement C3 measurement	4	1	0.000582	0.001985	3.41	3.15E-02	0.283	immune_system
complement C4 measurement	4	1	0.000582	0.001985	3.41	3.15E-02	0.283	immune_system
hepatitis B infection	3	1	0.000349	0.001489	4.26	3.46E-02	0.290	immune_system
infant body height	5	2	0.000873	0.002481	2.84	3.35E-02	0.290	body_measurement
kidney disease	3	1	0.000349	0.001489	4.26	3.46E-02	0.290	kidney
mean corpuscular hemoglobin	17	10	0.005124	0.008437	1.65	3.44E-02	0.290	blood
sporadic amyotrophic lateral sclerosis	18	11	0.005532	0.008933	1.61	3.54E-02	0.292	nervous_system
multiple myeloma	10	5	0.002562	0.004963	1.94	3.79E-02	0.307	cancer
conduct disorder	13	8	0.003727	0.006452	1.73	4.27E-02	0.329	brain_disease
narcolepsy	5	2	0.000932	0.002481	2.66	4.22E-02	0.329	brain_disease
short-term memory	4	1	0.000641	0.001985	3.10	4.21E-02	0.329	brain
atopic eczema	7	3	0.001630	0.003474	2.13	4.99E-02	0.349	immune_system
biliary atresia	2	0	0.000175	0.000993	5.68	4.92E-02	0.349	liver
eye measurement	6	3	0.001281	0.002978	2.32	4.76E-02	0.349	eye
nt-probnp measurement	2	0	0.000175	0.000993	5.68	4.92E-02	0.349	cardiovascular
response to bevacizumab	2	0	0.000175	0.000993	5.68	4.92E-02	0.349	drug_response
tooth agenesis	2	0	0.000175	0.000993	5.68	4.92E-02	0.349	dental

Table S10

Aggregate Grammatical Class	GWAS Ontology Term	Observed Peak Count	Expected Peak Count	Expected Frequency	Observed Frequency	Freq FC	Raw P-Value	Adjusted P-Value (fdr)	Category
MCKG	crohn's disease	25	11	0.012287	0.028027	2.28	1.69E-04	0.014	Immune System
	esophageal squamous cell carcinoma	5	1	0.000582	0.005605	9.63	2.03E-04	0.014	Cancer
	inflammatory bowel disease	19	7	0.007628	0.021300	2.79	8.39E-05	0.011	Immune System
	multiple sclerosis	23	11	0.012170	0.025785	2.12	8.11E-04	0.041	Immune System
	response to irinotecan	5	0	0.000524	0.005605	10.70	1.25E-04	0.013	drug_response
	schizophrenia	37	18	0.020439	0.041480	2.03	6.06E-05	0.011	brain_disease
	seasonal allergic rhinitis	10	2	0.002155	0.011211	5.20	3.24E-05	0.011	Immune System
	ulcerative colitis	17	7	0.007628	0.019058	2.50	6.70E-04	0.039	Immune System
Three cell	blood metabolite measurement	19	6	0.012694	0.040860	3.22	1.18E-05	1.74E-03	Blood
	erythrocyte measurement	9	1	0.002620	0.019355	7.39	5.17E-06	1.52E-03	Blood
	inflammatory bowel disease	13	4	0.007628	0.027957	3.66	7.91E-05	5.64E-03	Immune System
	mean corpuscular hemoglobin	9	2	0.005124	0.019355	3.78	7.86E-04	0.039	Blood
	metabolic syndrome	10	2	0.003901	0.021505	5.51	1.94E-05	1.90E-03	other_metabolic_disorder
	psoriasis	6	1	0.002387	0.012903	5.40	9.93E-04	0.042	Immune System
	systemic scleroderma	7	1	0.002213	0.015054	6.80	9.59E-05	5.64E-03	Immune System
Two cell miixed	immune system disease	4	0	0.001165	0.017391	14.93	1.69E-04	0.035	Immune System
	self reported educational attainment	6	1	0.004076	0.026087	6.40	4.06E-04	0.042	brain
Human specific	anti-Heliobacter pylori serum IgG measurement	3	0	0.000116	0.009063	77.82	9.19E-06	2.50E-03	Immune System
	erythrocyte measurement	6	1	0.002620	0.018127	6.92	2.74E-04	0.025	Blood
	platelet count	8	2	0.005357	0.024169	4.51	4.84E-04	0.033	Blood
	systemic lupus erythematosus	11	3	0.007861	0.033233	4.23	7.91E-05	0.011	Immune System

Table S11

Grammatical Class	GWAS Ontology Term	Observed Peak Count	Expected Peak Count	Expected Frequency	Observed Frequency	Freq FC	Raw P-Value	Adjusted P-Value (fdr)	Category
M	colorectal cancer	1	0	0.004950	0.333333	67.35	0.015	0.022	cancer
	immune system disease	1	0	0.001165	0.333333	286.22	3.49E-03	0.022	immune_system
	mean corpuscular hemoglobin	1	0	0.005124	0.333333	65.05	0.015	0.022	blood
	mean corpuscular volume	1	0	0.004134	0.333333	80.62	0.012	0.022	blood
	platelet count	1	0	0.005357	0.333333	62.22	0.016	0.022	blood
C	mean platelet volume	1	0	0.003086	0.333333	108.01	9.23E-03	0.028	blood
K	--	--	--	--	--	--	--	--	--
G	alcohol drinking	3	0	0.004542	0.060000	13.21	1.57E-03	0.037	behavior
	anti-Heliobacter pylori serum IgG measurement	3	0	0.000116	0.060000	515.19	3.08E-08	2.93E-06	immune_system
	celiac disease	3	0	0.003901	0.060000	15.38	1.01E-03	0.032	gut
	seasonal allergic rhinitis	3	0	0.002155	0.060000	27.85	1.82E-04	8.63E-03	immune_system
KG	total blood protein measurement	2	0	0.000408	0.032787	80.44	2.99E-04	0.044	blood
CG	--	--	--	--	--	--	--	--	--
CK	immune system disease	4	0	0.001165	0.017391	14.93	1.69E-04	0.035	immune_system
	self reported educational attainment	6	1	0.004076	0.026087	6.40	4.06E-04	0.042	brain
MC	--	--	--	--	--	--	--	--	--
MK	--	--	--	--	--	--	--	--	--
CKG	blood metabolite measurement	12	4	0.012694	0.038339	3.02	7.85E-04	0.032	blood
	colorectal cancer	7	2	0.004950	0.022364	4.52	1.07E-03	0.037	cancer
	crohn's disease	13	4	0.012287	0.041534	3.38	1.66E-04	0.010	immune_system
	erythrocyte measurement	7	1	0.002620	0.022364	8.53	2.30E-05	3.17E-03	blood
	immune system disease	4	0	0.001165	0.012780	10.97	5.42E-04	0.026	immune_system
	inflammatory bowel disease	11	2	0.007628	0.035144	4.61	3.69E-05	3.17E-03	immune_system
	mean corpuscular hemoglobin	7	2	0.005124	0.022364	4.36	1.30E-03	0.040	blood
	metabolic syndrome	8	1	0.003901	0.025559	6.55	3.91E-05	3.17E-03	other_metabolic_disorder
MCG	--	--	--	--	--	--	--	--	--
MCK	--	--	--	--	--	--	--	--	--
MCKG	crohn's disease	25	11	0.012287	0.028027	2.28	1.69E-04	0.014	immune_system
	esophageal squamous cell carcinoma	5	1	0.000582	0.005605	9.63	2.03E-04	0.014	cancer
	inflammatory bowel disease	19	7	0.007628	0.021300	2.79	8.39E-05	0.011	immune_system
	multiple sclerosis	23	11	0.012170	0.025785	2.12	8.11E-04	0.041	immune_system
	response to irinotecan	5	0	0.000524	0.005605	10.70	1.25E-04	0.013	drug_response
	schizophrenia	37	18	0.020439	0.041480	2.03	6.06E-05	0.011	brain_disease
	seasonal allergic rhinitis	10	2	0.002155	0.011211	5.20	3.24E-05	0.011	immune_system
	ulcerative colitis	17	7	0.007628	0.019058	2.50	6.70E-04	0.039	immune_system