Appendix

Arpita Joshi ✉ 🏠
The Scripps Research Institute, San Diego, USA

A Details of Select Metrics and The Top 10,000 ranked DHSs

- **Entropy:** Claude. E. Shannon, in his seminal work [4] described a communication model with a metric for the amount of information transmitted over a channel. It associates the probability of an event with the surprise associated with its occurrence as the logarithm of the inverse of the probability. For the genomic regions, we define the probability of a region (p) as the fraction of biosamples that register their presence in the associated genomic region. In the context of DHS data, this amounts to the fraction of calls (or the fraction of biosamples) that register their significant presence in the corresponding DHS.

  $$Entropy = -p * \log(p) \tag{1}$$

- **Average Normalized Signal:** Apart from the categorical matrix that presents 'call' (described in main text) biosamples, the actual signal levels that associate DNAase-I accessibility to a DHS are also available as a part of [2]. The signal levels associated with a DHS (or a genomic region in general) are not always comparable. With this metric we quantify the amount of information contained in a region by first normalizing the signal levels of only the biosamples that contribute a "call" associated with each DHS so that they lie between 0 and 1, and then obtaining the mean of these normalized signal values.
- **Mean TF-IDF:** TF-IDF (term frequency, inverse document frequency) is an information metric widely used in NLP (Natural Language Processing) [3] to evaluate the importance of tokens as they appear in various documents. We devised a variant of this metric for the DHS data, wherein we used the NMF annotations as documents. The term frequency is defined as the fraction of samples that have strong representation in a document/NMF component and have significant signal values in the genomic region. On the other hand, inverse document frequency is simply the logarithm of the reciprocal of the fraction of NMF components the genomic region has a presence in. This results in as many TF-IDF scores for each region as there are NMF components. We use the mean of all of these as the metric of information. The metric is useful in sampling genomics regions that are most represented in a given NMF component. However, the scores are less comparable across labels (NMF components in this case), which is why we resort to the concordance metric.

We created the heatmaps for a subset (top 10,000) of the DHSs that are high ranking according to a metric. The heatmaps were constructed using the actual signal values for each DHS. We scaled the signal values for each DHS to only have values between 0 and 1. We then fed the transformed sub-matrix to Python's scipy.cluster hierarchy function and used the ward linkage to obtain clusters within these DHSs. The columns are the top 10,000 highest ranked DHSs for each metric and the rows are the 733 samples ordered according to their respective NMF components. The order of components from top to bottom are: Placental/Trophoblast, Lymphoid, Myeloid/erythroid, Cardiac, Musculoskeletal, Vascular/endothelial, Primitive/embryonic, Neural, Digestive, Stromal A, Stromal B, Renal/cancer, Cancer/epithelial, Pulmonary development, Organ development.

Fig. 1 shows that entropy is maximum for DHSs that have calls in the range of 200-300 and Fig. 2 (a) has maximum signal for DHSs clustered towards the left of the figure, which largely belong to the Stromal-A or Stromal-B NMF component which have calls in the same

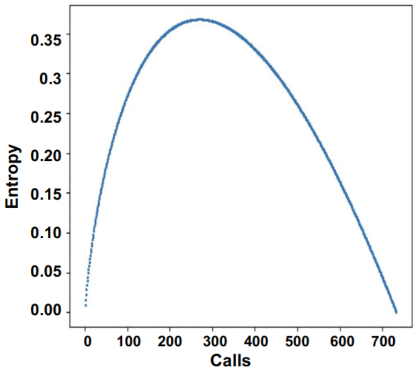


Figure 1 Entropy with the number of calls in a DHS

range. On the other hand, the mean cosine similarity metric rewards the regions that have calls in highly similar samples and hence the clustered DHSs for this metric, Fig. 2 (d), have much fewer calls and are concentrated in compact regions representing their respective NMF components. In summary, each of the unbiased metrics captures a unique property of DHS regions and combining them to select important ones makes more sense than developing one catch-all metric.

B Details of Co-ranking and the Concoradance Metric

- Fig. 3 shows the pictorial description of the computation of the Concordance Metric.
- As delineated in the Methods, we performed hierarchical clustering on the top 10,000 DHSs. Fig.4 displays the heatmap with 9 clusters obtained by ranking according to the Concordance Metric, and the figure underneath is the signal distribution of one representative DHS from each cluster. The signal representation from each component is obtained as a screenshot from the Index Browser: https://index.altius.org/. The representative DHS is chosen as the one nearest to the centroid of the cluster. Fig.5 shows the same for the co-ranking of the Mean Cosine Similarity and the Signal to Noise Ratio.

C ARCHS4: Ranking Metrics and Gene Expression Across Tissue Types

- **The Signal Metric:** To approximate the SNR metric for gene expression, we modified the metric to get the 95 percentile expression value for each gene, making the metric more robust to noise errors from both very small and very large number of readcounts, given the high dimensionality of the data. The expression values from the low expressed biosamples for each gene present a near constant noise level for each gene, likewise the highest expression values are prone to errors from the Kallisto aligner[1] leading to misjudgement in the relative expression values of biosamples within a gene. The 95 percentile expression value hence is a better measure of signal contained in a gene. Fig. **??** (a) through (d) show the kernel density estimates for the top genes, as identified by this metric.
- **Mean Cosine Similarity:** Again, we identified 100 biosamples around the 95 percentile signal/expression value for each gene, and computed the mean cosine similarity among them as described earlier. Fig. **??** (e) through (h) show the kernel density estimates for the highest ranked genes by this metric.

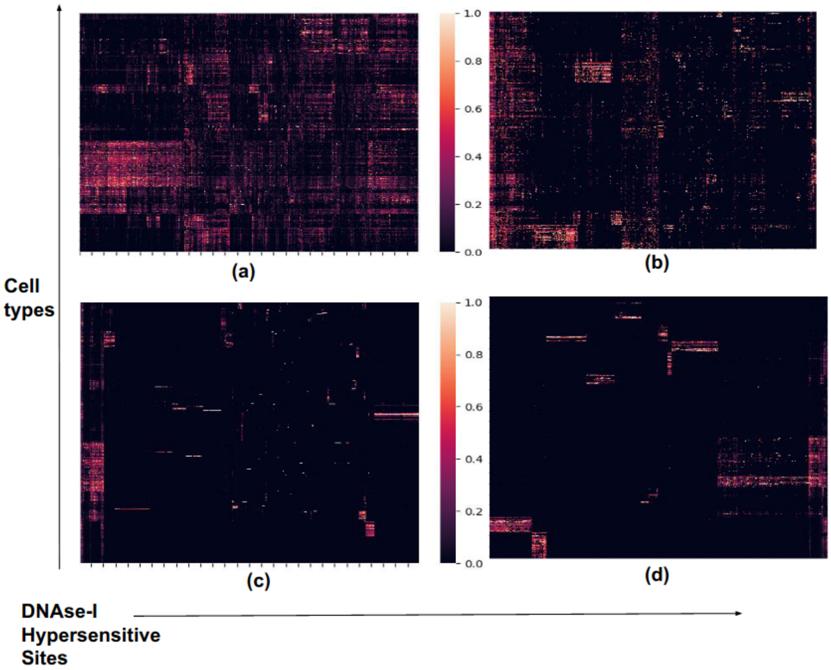


Figure 2 The clustered DHSs for each of the four unbiased ranking metrics (a) Entropy, (b) Average Normalized Signal, (c) Signal to Noise Ratio, (d) Mean Cosine Similarity

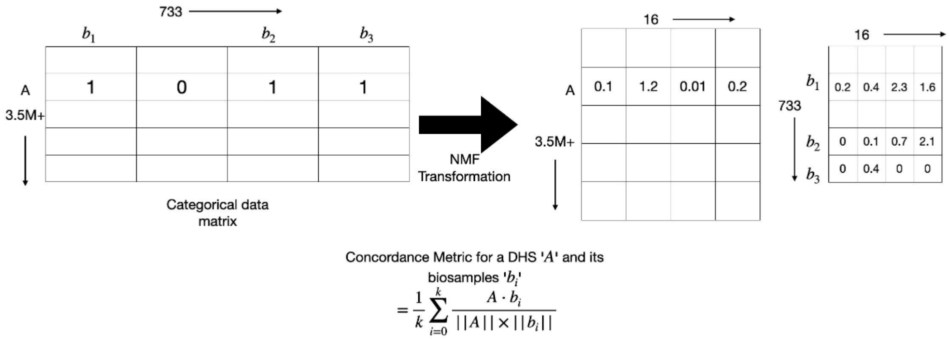


Figure 3 The data matrix of dimensionality 3.5M+ by 733 is transformed into NMF projections of dimensionality 3.5M+ by 16 and 733 by 16. The Concordance metric is then computed as the mean of cosine similarity between a DHS and all of its constituent biosamples.

4 Appendix

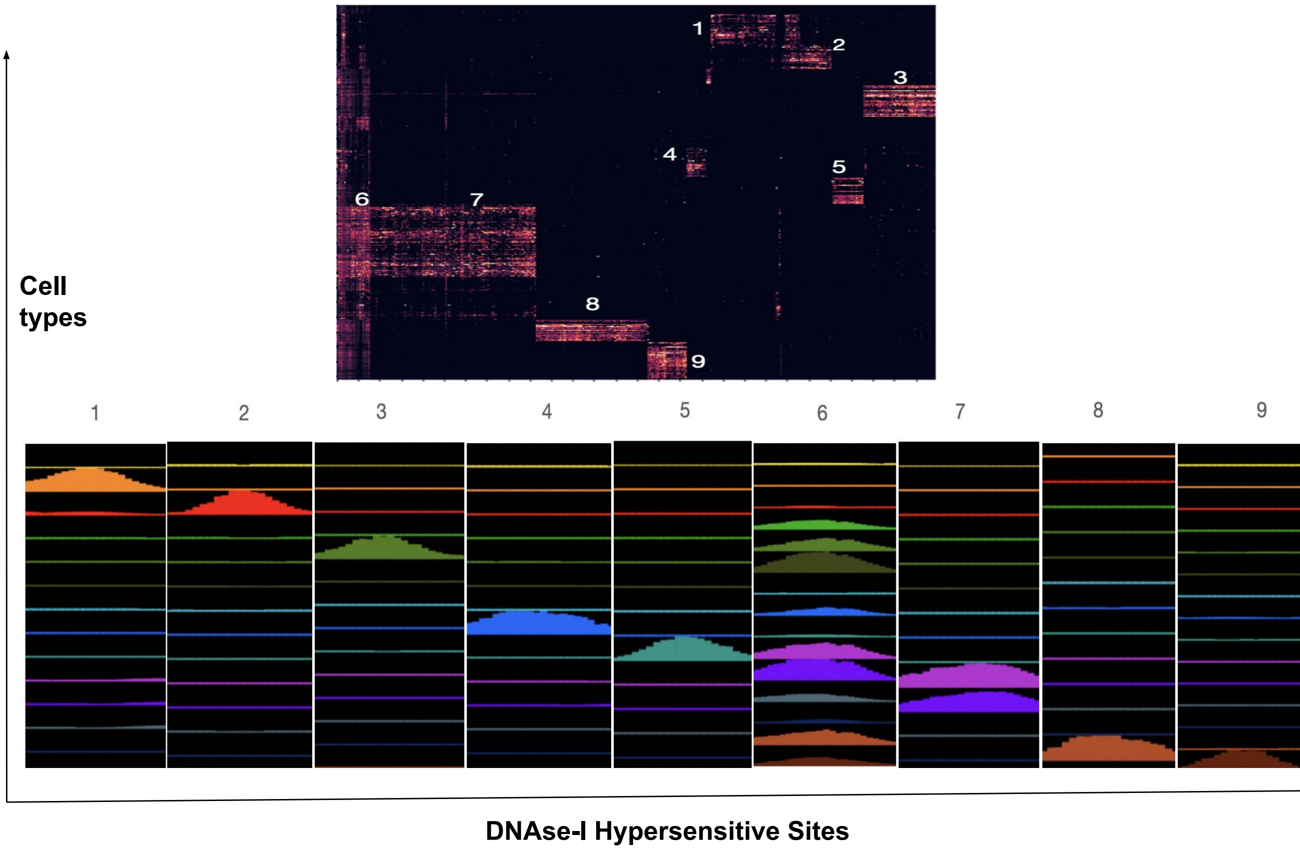


Figure 4 The cluster of DHSs by Concordance Metric and the representative DHS from each cluster.

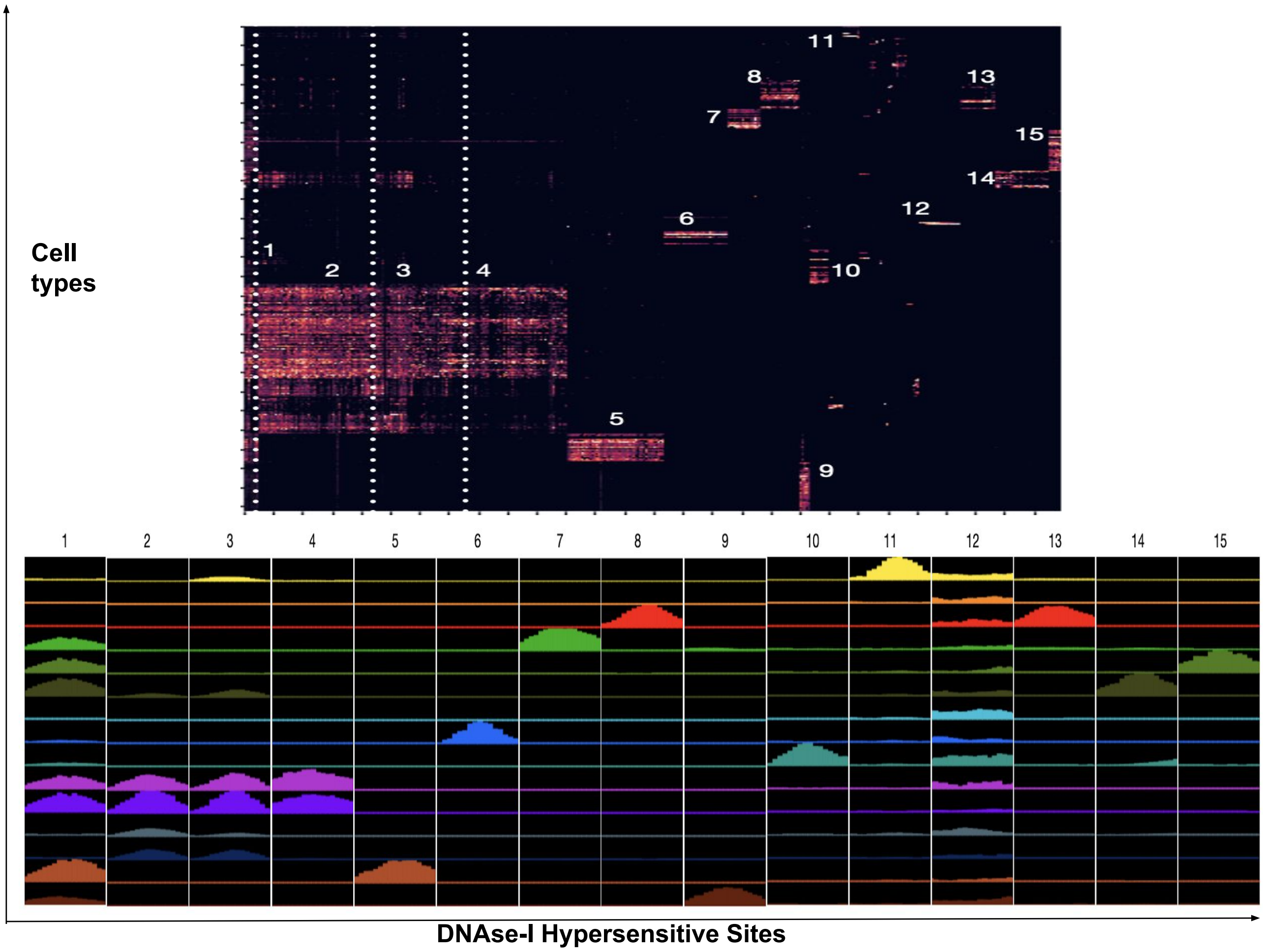
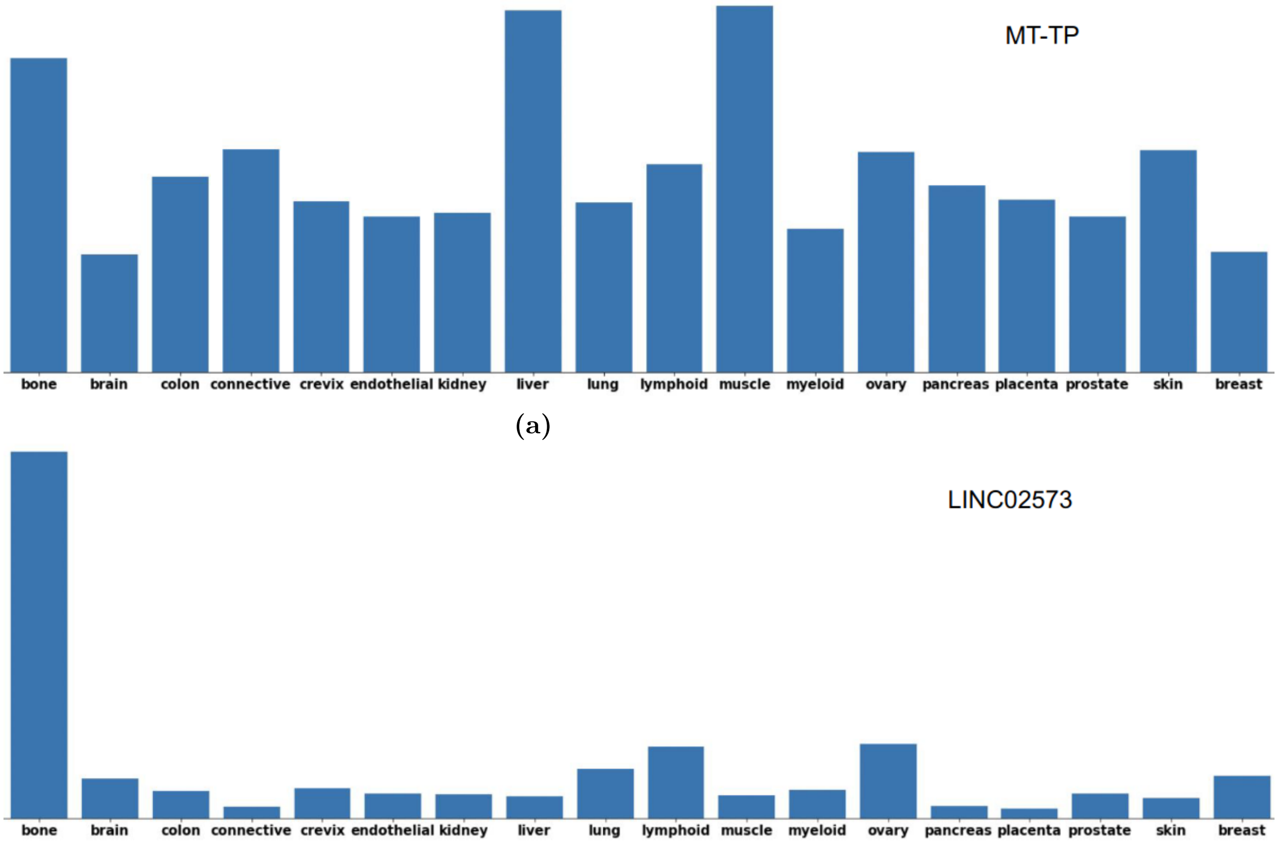


Figure 5 The heatmap and representative DHSs from the clusters on heatmap of the top 10,000 best ranked DHSs by the co-ranking of Mean Cosine Similarity and Signal to Noise Ratio



(a)

(b)

Figure 6 Relative expression in biosamples of corresponding tissue types in genes: (a) MT-TP: sampled as one of the highest ranked genes by SNR, (b) LINC02573: sampled as one of the highest ranked genes by MCS

As corroboratory evidence for claims about the traits of genes selected by SNR and MCS metrics, we present the expression across tissue types for both the genes *MT-TP* and *LINC02573* in Fig. 6.

D 150 Highest Ranked 25kb DHS Windows

Please find the summary statistics of the top 150 highest ranked 25kb DHS windows in the Table 1. The first three columns describe the genomic region, the fourth column is the number of DHSs found in the region, the fifth column is the p-value obtained as a result of CLT (Central Limit Theorem), signifying the importance of the region in terms of the mean SNR and MCS scores of its constituent DHSs, the sixth column is the maximum of the 16 enrichment scores corresponding to the 16 NMF components/cell-types, and the last column is the maximally enriched cell-type.

Table 1 Summary statistics of the top 150 25kb windows from DHS data

seqname	start	end	no_of_dhs	clt_pvalue	max_enrich	winner_nmf_component
chr4	173514800	173539800	94	0.000e+0	4.901e+0	Cardiac
chr17	48575800	48600800	100	0.000e+0	4.677e+0	Pulmonary devel.
chr5	93578200	93603200	95	0.000e+0	4.769e+0	Organ devel. / renal
chr17	48541000	48566000	101	2.220e-16	5.391e+0	Pulmonary devel.
chr7	27093600	27118600	98	0.000e+0	4.354e+0	Organ devel. / renal
chr15	37090600	37115600	93	0.000e+0	4.190e+0	Cardiac
chr10	117530600	117555600	101	0.000e+0	4.354e+0	Organ devel. / renal
chr7	27178600	27203600	95	0.000e+0	4.217e+0	Organ devel. / renal
chrY	11314400	11339400	61	3.331e-16	4.387e+0	Vascular / endothelial
chr11	119358800	119383800	87	0.000e+0	4.053e+0	Cardiac
chr9	136842200	136867200	95	0.000e+0	3.971e+0	Digestive
chr20	58699600	58724600	88	0.000e+0	4.033e+0	Musculoskeletal
chr2	104848800	104873800	94	0.000e+0	4.010e+0	Organ devel. / renal
chr17	48602000	48627000	88	2.220e-15	4.310e+0	Organ devel. / renal
chr10	21514800	21539800	98	0.000e+0	4.053e+0	Cardiac
chr4	13524600	13549600	93	1.110e-16	4.065e+0	Organ devel. / renal
chr15	96332600	96357600	99	0.000e+0	3.954e+0	Organ devel. / renal
chrY	11289400	11314400	78	2.476e-14	4.465e+0	Vascular / endothelial
chr14	77022200	77047200	104	0.000e+0	3.901e+0	Cardiac
chr2	176142800	176167800	89	1.213e-13	4.667e+0	Organ devel. / renal
chr7	35672800	35697800	71	1.332e-15	4.048e+0	Lymphoid
chr19	42267600	42292600	107	0.000e+0	3.863e+0	Musculoskeletal
chr9	129080200	129105200	73	0.000e+0	3.820e+0	Cancer / epithelial
chr12	68748000	68773000	38	6.035e-13	4.899e+0	Stromal A
chr9	14306600	14331600	87	0.000e+0	3.818e+0	Cardiac
chr7	27125000	27150000	98	0.000e+0	3.841e+0	Pulmonary devel.
chr12	92124200	92149200	93	4.171e-13	4.486e+0	Cardiac
chr18	22158200	22183200	84	8.349e-13	4.638e+0	Cardiac
chr8	11685400	11710400	88	1.122e-12	4.685e+0	Cardiac
chr14	105467200	105492200	95	5.757e-13	4.375e+0	Cardiac
chr6	44216000	44241000	92	1.221e-15	3.901e+0	Cardiac
chr7	38055000	38080000	52	5.054e-13	4.275e+0	Cancer / epithelial
chr12	53971600	53996600	87	1.965e-13	4.065e+0	Organ devel. / renal
chr6	123403400	123428400	54	1.308e-12	4.195e+0	Cancer / epithelial
chr3	42012600	42037600	83	2.607e-12	4.316e+0	Cardiac
chr7	44200600	44225600	84	3.331e-16	3.784e+0	Placental / trophoblast
chr6	1601800	1626800	88	0.000e+0	3.702e+0	Organ devel. / renal
chr6	1373000	1398000	77	1.887e-14	3.841e+0	Pulmonary devel.
chr2	219625000	219650000	78	1.110e-16	3.731e+0	Cardiac
chr7	134084600	134109600	66	1.623e-13	3.907e+0	Digestive
chrX	72180400	72205400	64	7.438e-14	3.851e+0	Lymphoid
chr12	53055200	53080200	90	1.033e-14	3.817e+0	Musculoskeletal
chr4	173489800	173514800	80	5.050e-12	4.316e+0	Cardiac
chr20	63696200	63721200	79	4.441e-15	3.779e+0	Lymphoid

8 Appendix

chr6	137322600	137347600	61	2.286e-12	4.066e+0	Cancer / epithelial
chr1	3055200	3080200	90	1.758e-12	4.010e+0	Organ devel. / renal
chr7	1839000	1864000	93	2.951e-12	4.065e+0	Organ devel. / renal
chr1	145977200	146002200	91	1.367e-11	4.375e+0	Cardiac
chr4	2238600	2263600	96	0.000e+0	3.632e+0	Organ devel. / renal
chr2	66423000	66448000	103	0.000e+0	3.632e+0	Organ devel. / renal
chrY	56839400	56864400	70	0.000e+0	3.617e+0	Digestive
chrX	321800	346800	52	8.882e-16	3.662e+0	Placental / trophoblast
chr5	180995800	181020800	45	3.870e-12	3.938e+0	Myeloid / erythroid
chr8	139660800	139685800	53	4.397e-12	3.940e+0	Pulmonary devel.
chrY	7789400	7814400	60	4.773e-13	3.816e+0	Lymphoid
chr9	136606600	136631600	87	3.905e-11	4.264e+0	Organ devel. / renal
chr10	129953600	129978600	88	3.886e-15	3.670e+0	Musculoskeletal
chr7	51646600	51671600	52	4.759e-12	3.924e+0	Cancer / epithelial
chr22	31080400	31105400	94	1.110e-13	3.721e+0	Musculoskeletal
chr8	39532200	39557200	52	2.552e-12	3.840e+0	Digestive
chr11	5262600	5287600	43	1.173e-13	3.724e+0	Myeloid / erythroid
chr5	177979000	178004000	73	2.220e-16	3.623e+0	Lymphoid
chr18	79382000	79407000	89	1.443e-15	3.632e+0	Organ devel. / renal
chr2	176116000	176141000	97	0.000e+0	3.558e+0	Organ devel. / renal
chr7	23787600	23812600	56	4.927e-11	4.066e+0	Cancer / epithelial
chr15	63041000	63066000	91	2.188e-10	4.663e+0	Renal / cancer
chr4	144115400	144140400	43	5.725e-12	3.835e+0	Myeloid / erythroid
chr2	36354400	36379400	97	1.839e-10	4.465e+0	Vascular / endothelial
chr1	156727400	156752400	85	1.212e-12	3.731e+0	Cardiac
chr2	176167800	176192800	86	1.216e-11	3.833e+0	Organ devel. / renal
chr4	54217600	54242600	75	3.331e-16	3.571e+0	Tissue invariant
chr16	30948000	30973000	68	2.937e-12	3.766e+0	Cancer / epithelial
chrX	74559400	74584400	45	1.787e-14	3.623e+0	Lymphoid
chr3	196586000	196611000	68	6.158e-11	3.972e+0	Cancer / epithelial
chr19	13148000	13173000	91	8.882e-16	3.571e+0	Tissue invariant
chr7	39091400	39116400	57	4.209e-10	4.458e+0	Cancer / epithelial
chr20	5095800	5120800	68	5.181e-13	3.651e+0	Cancer / epithelial
chr11	117861400	117886400	78	3.563e-12	3.731e+0	Cardiac
chr8	143904000	143929000	76	5.878e-11	3.907e+0	Musculoskeletal
chr8	127056600	127081600	52	3.123e-14	3.604e+0	Myeloid / erythroid
chr17	8140000	8165000	90	0.000e+0	3.507e+0	Musculoskeletal
chr12	54030200	54055200	88	7.481e-10	4.667e+0	Organ devel. / renal
chr5	139741200	139766200	91	8.882e-15	3.558e+0	Organ devel. / renal
chr2	90377400	90402400	62	1.526e-12	3.664e+0	Lymphoid
chr1	27553000	27578000	91	8.166e-12	3.731e+0	Cardiac
chr8	20348200	20373200	72	3.236e-13	3.623e+0	Lymphoid
chr6	27806800	27831800	86	0.000e+0	3.472e+0	Myeloid / erythroid
chr6	27124600	27149600	86	3.053e-14	3.539e+0	Myeloid / erythroid
chr3	52229400	52254400	91	6.761e-14	3.558e+0	Organ devel. / renal
chr4	10220200	10245200	66	2.805e-11	3.769e+0	Digestive
chr19	35724400	35749400	74	7.755e-11	3.822e+0	Placental / trophoblast
chrX	132623800	132648800	55	2.894e-11	3.766e+0	Cancer / epithelial
chr8	30382200	30407200	49	6.694e-11	3.818e+0	Cardiac
chr3	195502000	195527000	68	4.582e-11	3.779e+0	Lymphoid
chr2	28389400	28414400	92	1.074e-9	4.311e+0	Renal / cancer

chr7	135297000	135322000	74	7.550e-15	3.525e+0	Cancer / epithelial
chr3	52388400	52413400	81	8.185e-10	4.185e+0	Musculoskeletal
chr1	43430000	43455000	92	2.109e-15	3.507e+0	Musculoskeletal
chrY	56814400	56839400	60	0.000e+0	3.447e+0	Digestive
chr17	82041600	82066600	76	1.262e-10	3.818e+0	Cardiac
chr3	194253800	194278800	89	2.279e-13	3.539e+0	Vascular / endothelial
chr20	58888000	58913000	86	2.384e-12	3.618e+0	Musculoskeletal
chr22	36363800	36388800	84	2.115e-13	3.538e+0	Cardiac
chr6	44245600	44270600	86	2.254e-14	3.494e+0	Lymphoid
chr9	134199800	134224800	70	1.475e-9	4.168e+0	Organ devel. / renal
chr9	38045600	38070600	89	0.000e+0	3.431e+0	Cardiac
chr6	163404400	163429400	59	1.255e-11	3.638e+0	Cardiac
chr3	129603400	129628400	81	9.859e-14	3.507e+0	Musculoskeletal
chr10	75393800	75418800	89	5.307e-14	3.493e+0	Pulmonary devel.
chr20	1555200	1580200	77	7.726e-11	3.703e+0	Placental / trophoblast
chr13	113642800	113667800	71	1.845e-11	3.632e+0	Organ devel. / renal
chr9	35055800	35080800	59	0.000e+0	3.411e+0	Tissue invariant
chr6	112064000	112089000	55	9.733e-10	3.924e+0	Cancer / epithelial
chr20	54909400	54934400	37	6.864e-10	3.873e+0	Cancer / epithelial
chr7	38595200	38620200	48	9.882e-10	3.924e+0	Cancer / epithelial
chr22	23514800	23539800	84	2.089e-11	3.632e+0	Organ devel. / renal
chr2	179215800	179240800	77	3.061e-9	4.289e+0	Myeloid / erythroid
chr2	41975000	42000000	82	7.858e-10	3.895e+0	Organ devel. / renal
chr4	157430600	157455600	48	2.189e-10	3.742e+0	Lymphoid
chr5	2289800	2314800	67	1.901e-9	4.033e+0	Pulmonary devel.
chrX	55717000	55742000	54	7.691e-12	3.581e+0	Lymphoid
chr17	7833800	7858800	100	0.000e+0	3.411e+0	Tissue invariant
chr9	35689000	35714000	61	1.430e-13	3.481e+0	Placental / trophoblast
chr5	132477200	132502200	90	7.550e-15	3.448e+0	Lymphoid
chr8	46491200	46516200	35	1.200e-11	3.581e+0	Lymphoid
chr1	156113000	156138000	82	6.007e-10	3.812e+0	Stromal A
chr9	137335800	137360800	64	1.718e-10	3.703e+0	Lymphoid
chr21	10780600	10805600	38	1.903e-10	3.703e+0	Lymphoid
chrX	65997600	66022600	58	6.049e-9	4.438e+0	Myeloid / erythroid
chr3	191024400	191049400	56	1.268e-10	3.651e+0	Cancer / epithelial
chr6	41634400	41659400	88	1.454e-11	3.564e+0	Musculoskeletal
chr12	57510800	57535800	78	1.443e-14	3.432e+0	Placental / trophoblast
chr8	143830200	143855200	61	6.380e-10	3.770e+0	Musculoskeletal
chr17	2048000	2073000	105	1.110e-16	3.397e+0	Organ devel. / renal
chr4	10180000	10205000	101	7.879e-9	4.438e+0	Myeloid / erythroid
chr5	176365000	176390000	68	1.184e-12	3.481e+0	Placental / trophoblast
chr15	96307600	96332600	97	1.405e-9	3.833e+0	Organ devel. / renal
chr8	144265400	144290400	71	3.713e-9	4.010e+0	Organ devel. / renal
chr1	234956800	234981800	94	7.793e-11	3.619e+0	Stromal A
chr5	91360800	91385800	75	3.605e-11	3.571e+0	Tissue invariant
chr7	97002800	97027800	93	0.000e+0	3.389e+0	Neural
chr14	37582200	37607200	82	3.095e-13	3.447e+0	Digestive
chr22	38482400	38507400	75	7.072e-14	3.431e+0	Cardiac
chr12	34696400	34721400	31	4.683e-9	3.924e+0	Cancer / epithelial
chr8	143600800	143625800	81	0.000e+0	3.387e+0	Musculoskeletal
chr18	48928400	48953400	101	0.000e+0	3.360e+0	Renal / cancer
chr5	88875200	88900200	79	4.328e-10	3.638e+0	Cardiac
chr22	36220400	36245400	81	8.218e-12	3.481e+0	Placental / trophoblast
chr5	179816600	179841600	82	9.992e-16	3.387e+0	Vascular / endothelial
chr9	20600200	20625200	68	5.932e-13	3.431e+0	Cardiac

10 Appendix

References

1 Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(5):525–527, May 2016.

2 Wouter Meuleman, Alexander Muratov, Eric Rynes, Jessica Halow, Kristen Lee, Daniel Bates, Morgan Diegel, Douglas Dunn, Fidencio Neri, Athanasios Teodosiadis, Alex Reynolds, Eric Haugen, Jemma Nelson, Audra Johnson, Mark Frerker, Michael Buckley, Richard Sandstrom, Jeff Vierstra, Rajinder Kaul, and John Stamatoyannopoulos. Index and biological spectrum of human DNase I hypersensitive sites. *Nature*, 584(7820):244–251, August 2020.

3 Juan Ramos. Using TF-IDF to determine word relevance in document queries. `https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b3bf6373ff41a115197cb5b30e57830c16130c2c`. Accessed: 2023-1-17.

4 C E Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.