nature methods

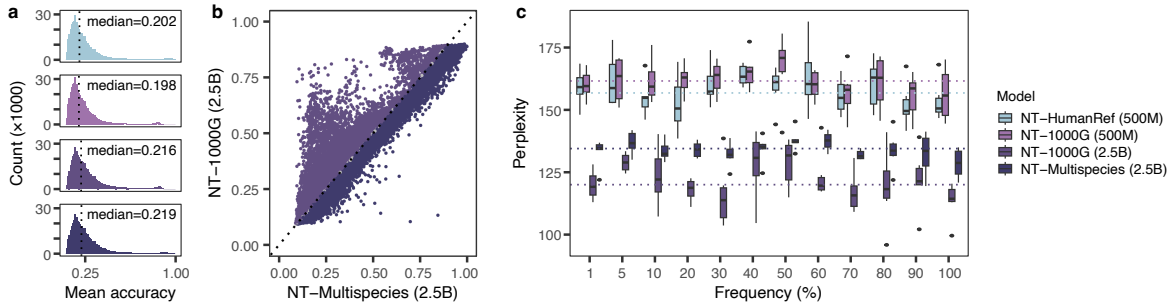
Supplementary information

https://doi.org/10.1038/s41592-024-02523-z

Nucleotide Transformer: building and evaluating robust foundation models for human genomics

In the format provided by the authors and unedited

A Supplementary Note: Nucleotide Transformer learned to reconstruct genetic variants



Supplementary Note Figure 1: The Nucleotide Transformer models acquired knowledge about genetic variations and genomic elements. **a)** Mean reconstruction accuracy of 6kb sequences across transformer models. **b)** Comparison of reconstruction accuracies between the 1000G 2.5B and Multispecies 2.5B models. **c)** Reconstruction perplexity across allele frequencies. Perplexity values shown per frequency bin are based on 6 meta-populations from the Human Genome Diversity Project (n=6 per boxplot). The boxplots mark the median, upper and lower quartiles, and 1.5× interquartile range (whiskers).

To investigate the advantages of increasing the number of parameters in the models and enhancing the genomic diversity of the training datasets, we evaluated the models' ability to reconstruct masked nucleotides. Specifically, we divided the human reference genome into non-overlapping 6kb sequences, tokenized each sequence into 6-mers, randomly masked a certain number of tokens, and then calculated the proportion of tokens that were accurately reconstructed (Supplementary Note Fig. 1a; Supplementary Fig. 4, Methods). We observed that the reconstruction accuracy of the Human reference 500M model exhibited higher median accuracy compared to the 1000G 500M model (median=0.202 versus 0.198, P<2.2e-16, two-sided Wilcoxon rank sum test). However, the accuracy of this model was lower than that achieved by the 1000G 2.5B model (median=0.216; P<2.2e-16, two-sided Wilcoxon rank sum test) and the 2.5B Multispecies model (median=0.219; P<2.2e-16, two-sided Wilcoxon rank sum test) (Supplementary Fig. 4), highlighting the impact of simultaneously increasing the model size and diversifying the dataset. Interestingly, while the Multispecies 2.5B model exhibited the highest overall reconstruction accuracy, specific sequences demonstrated significantly higher reconstruction accuracy for the Human 1000G 2.5B model (Supplementary Note Fig. 1b). This likely indicates that certain characteristics of human sequences were better captured and learned by the latter model. For example, it is possible that genomic annotations and molecular phenotypes controlled by elements overlapping these sequences might be better predicted by it.

To gain a deeper understanding of how transformer language models represent variant sites in DNA and their capabilities in imputing such variants, we then focused on polymorphisms identified in human samples from diverse populations. In order to assess whether our models can generalize and reconstruct variants in unseen human genome sequences, we evaluated their sequence reconstruction effectiveness by recording model perplexity scores across single nucleotide polymorphisms (SNPs) occurring at frequencies ranging from 1% to 100% (Methods). To account for the influence of genomic background and genetic structure on the mutation reconstruction, we considered an independent dataset of genetically diverse human genomes, originating from 6 different meta-populations [**?**] (see Methods). Across varying SNP frequencies, we consistently observed that the 2.5B parameter models, with median perplexity across populations ranging from 95.9±17.9 (2SD) to 145±9.3, outperformed their 500M counterparts, for which the perplexity values fluctuate between 138.5±8.4 and 185.4±9.5 (Supplementary Note Fig. 1c). Interestingly, the 1000G 2.5B model exhibited lower perplexity scores than the Multispecies 2.5B (median=121.5 versus 134.0, P<1.8e-12, two-sided Wilcoxon rank sum test) indicating that it effectively leveraged human genetic variability observed in the 1000G data to accurately reconstruct variants in unseen human genomes. We confirmed this result by analyzing the models' variant recon-

struction accuracy after excluding variants present in any individual from the 1000 Genomes Project, showing that the 1000G 2.5B model is leveraging relevant genomic features learned across diverse human sequences beyond simply allele frequencies (Supplementary Fig. 5). These results are in line with the observed reconstruction accuracies over human genome sequences, and confirm the impact of model size on performance (Supplementary Note Fig. 1b). Notably, and in contrast to typical genotype imputation methods where accuracy declines as variant frequency decreases [**?**], our models exhibited comparable performance across frequencies, suggesting their potential to enhance genotype imputation even for variants present at lower frequencies.

B Supplementary Tables

	NT-1000G (2.5B)	NT-Multispecies (2.5B)	NT-1000G (500M)	NT-HumanRef (500M)	NT-Multispecies-v2 (500M)	NT-Multispecies-v2 (250M)	NT-Multispecies-v2 (100M)	NT-Multispecies-v2 (50M)
alphabet	k-mers	k-mers	k-mers	k-mers	k-mers	k-mers	k-mers	k-mers
k_for_kmers	6	6	6	6	6	6	6	6
num_warmup_updates	16000	16000	16000	16000	16000	16000	16000	16000
warmup_init_lr	5E-05	5E-05	5E-05	5E-05	5E-05	5E-05	5E-05	5E-05
warmup_end_lr	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
training_set_proportion	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
masking_ratio	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
masking_prob	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
random_token_prob	0	0.1	0	0.1	0.1	0.1	0.1	0.1
alphabet_size	4105	4105	4105	4105	4105	4105	4105	4105
prepend_bos	True	True	True	True	True	True	True	True
append_eos	False	False	False	False	False	False	False	False
attention_heads	20	20	20	20	16	16	16	16
embed_dim	2560	2560	1280	1280	1024	768	512	512
ffn_embed_dim	10240	10240	5120	5120	4096	3072	2048	2048
num_layers	32	32	24	24	29	24	22	12
token_dropout	True	True	True	True	False	False	False	False
num_parameters	2547800585	2547800585	485729545	485729545	496245771	233545995	96839179	54855179
dataset	1000G	Multispecies	1000G	Human Reference	Multispecies	Multispecies	Multispecies	Multispecies

Supplementary Table 1: Models hyperparameters.

Population name	Population code	Number of individuals
African Ancestry SW	ASW	74
African Caribbean	ACB	116
Bengali	BEB	131
British	GBR	91
CEPH	CEU	179
Colombian	CLM	132
Dai Chinese	CDX	93
Esan	ESN	149
Finnish	FIN	99
Gambian Mandinka	GWD	178
Gujarati	GIH	103
Han Chinese	CHB	103
Iberian	IBS	157
Japanese	JPT	104
Kinh	KHV	2
Kinh Vietnamese	KHV	120
Luhya	LWK	99
Mende	MSL	99
Mexican Ancestry	MXL	97
Peruvian	PEL	122
Puerto Rican	PUR	139
Punjabi	PJL	146
Southern Han Chinese	CHS	163
Tamil	STU	114
Telugu	ITU	107
Toscani	TSI	107
Yoruba	YRI	178

Supplementary Table 2: Number of individuals per population in the 1000G dataset.

Class	Number of species	Number of nucleotides (B)
Bacteria	667	17.1
Fungi	46	2.3
Invertebrate	39	20.8
Protozoa	10	0.5
Mammalian Vertebrate	31	69.8
Other Vertebrate	57	63.4

Supplementary Table 3: Genomes in the multispecies dataset.

Class	Selected Species
Bacteria	Escherichia coli
Fungi	Saccharomyces cerevisiae
Invertebrate	Caenorhabditis elegans, Drosophila melanogaster
Protozoa	Plasmodium vivax, Plasmodium falciparum
Mammalian Vertebrate	Homo sapiens, Mus musculus, Rattus norvegicus
Other Vertebrate	Danio rerio, Xenopus tropicalis

Supplementary Table 4: Model organisms genomes in the multispecies dataset.

	Num train sequences	Num test sequences	Max sequence length in bp	Reported Metric	Organisms
H2AFZ	30000	3000	1000	MCC	Human
H3K27ac	30000	1616	1000	MCC	Human
H3K27me3	30000	3000	1000	MCC	Human
H3K36me3	30000	3000	1000	MCC	Human
H3K4me1	30000	3000	1000	MCC	Human
H3K4me2	30000	2138	1000	MCC	Human
H3K4me3	17468	776	1000	MCC	Human
H3K9ac	23274	1004	1000	MCC	Human
H3K9me3	27438	850	1000	MCC	Human
H4K20me1	30000	2270	1000	MCC	Human
Promoter all	30000	1584	300	MCC	Human
Promoter TATA	30000	1372	300	MCC	Human
Promoter non-TATA	5062	212	300	MCC	Human
Splice site all	30000	3000	600	MCC	Human
Splice donor	30000	3000	600	MCC	Human
Splice acceptor	30000	3000	600	MCC	Human
Enhancer	30000	3000	400	MCC	Human
Enhancer types	30000	3000	400	MCC	Human

Supplementary Table 5: Nucleotide Transformer downstream tasks metadata.

Dataset Model	H2AFZ	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me2
Raw Probe	0.302 (± 0.003)	0.229 (± 0.006)	0.379 (± 0.004)	0.36 (± 0.003)	0.287 (± 0.006)	0.36 (± 0.005)
BPNet (original)	**0.473 (± 0.009)**	0.296 (± 0.046)	0.543 (± 0.009)	0.548 (± 0.009)	0.436 (± 0.008)	0.427 (± 0.036)
BPNet (large)	**0.487 (± 0.014)**	0.214 (± 0.037)	0.551 (± 0.009)	**0.57 (± 0.009)**	**0.459 (± 0.012)**	0.427 (± 0.025)
BPNet (X-large)	0.475 (± 0.011)	0.16 (± 0.042)	0.542 (± 0.009)	0.562 (± 0.011)	0.457 (± 0.013)	0.341 (± 0.038)
NT-HumanRef (500M)	0.393 (± 0.01)	0.422 (± 0.01)	0.547 (± 0.007)	0.516 (± 0.01)	0.425 (± 0.014)	0.453 (± 0.01)
NT-1000G (500M)	0.404 (± 0.009)	0.437 (± 0.007)	0.535 (± 0.006)	0.509 (± 0.005)	**0.438 (± 0.012)**	0.515 (± 0.006)
NT-1000G (2.5B)	0.425 (± 0.01)	**0.457 (± 0.011)**	**0.572 (± 0.009)**	**0.579 (± 0.009)**	**0.443 (± 0.015)**	**0.531 (± 0.007)**
NT-Multispecies (2.5B)	0.422 (± 0.011)	**0.442 (± 0.019)**	0.552 (± 0.009)	**0.569 (± 0.011)**	0.427 (± 0.014)	0.512 (± 0.007)

Dataset Model	H3K4me3	H3K9ac	H3K9me3	H4K20me1	Enhancer	Enhancer types
Raw Probe	0.432 (± 0.008)	0.322 (± 0.013)	0.105 (± 0.01)	0.446 (± 0.006)	0.251 (± 0.003)	0.233 (± 0.004)
BPNet (original)	0.445 (± 0.047)	0.336 (± 0.034)	0.298 (± 0.03)	0.531 (± 0.025)	**0.488 (± 0.009)**	**0.449 (± 0.006)**
BPNet (large)	0.445 (± 0.049)	0.298 (± 0.033)	0.234 (± 0.037)	0.525 (± 0.038)	**0.492 (± 0.008)**	**0.454 (± 0.008)**
BPNet (X-large)	0.338 (± 0.076)	0.235 (± 0.06)	0.216 (± 0.085)	0.411 (± 0.054)	**0.501 (± 0.013)**	**0.455 (± 0.012)**
NT-HumanRef (500M)	**0.594 (± 0.012)**	0.458 (± 0.017)	0.28 (± 0.021)	0.608 (± 0.007)	0.455 (± 0.011)	0.423 (± 0.012)
NT-1000G (500M)	**0.592 (± 0.011)**	**0.523 (± 0.018)**	**0.381 (± 0.021)**	**0.622 (± 0.008)**	0.47 (± 0.008)	0.431 (± 0.005)
NT-1000G (2.5B)	**0.603 (± 0.018)**	**0.517 (± 0.014)**	**0.38 (± 0.031)**	**0.628 (± 0.012)**	0.479 (± 0.007)	**0.439 (± 0.009)**
NT-Multispecies (2.5B)	**0.604 (± 0.016)**	**0.539 (± 0.015)**	**0.396 (± 0.011)**	**0.611 (± 0.011)**	0.468 (± 0.009)	**0.446 (± 0.007)**

Dataset Model	Promoter all	Promoter non-TATA	Promoter TATA	Splice acceptor	Splice site all	Splice donor
Raw Probe	0.598 (± 0.002)	0.619 (± 0.004)	0.606 (± 0.007)	0.136 (± 0.04)	0.234 (± 0.003)	0.395 (± 0.025)
BPNet (original)	0.696 (± 0.026)	**0.717 (± 0.023)**	**0.848 (± 0.042)**	**0.859 (± 0.038)**	**0.793 (± 0.072)**	**0.92 (± 0.014)**
BPNet (large)	0.672 (± 0.023)	0.672 (± 0.043)	**0.826 (± 0.017)**	**0.925 (± 0.031)**	**0.865 (± 0.026)**	**0.93 (± 0.021)**
BPNet (X-large)	0.579 (± 0.087)	0.608 (± 0.061)	0.607 (± 0.064)	0.744 (± 0.01)	0.745 (± 0.016)	0.706 (± 0.009)
NT-HumanRef (500M)	0.703 (± 0.007)	**0.725 (± 0.011)**	0.662 (± 0.044)	0.485 (± 0.012)	0.414 (± 0.005)	0.495 (± 0.01)
NT-1000G (500M)	0.715 (± 0.011)	**0.732 (± 0.009)**	0.696 (± 0.018)	0.552 (± 0.006)	0.355 (± 0.01)	0.542 (± 0.01)
NT-1000G (2.5B)	**0.73 (± 0.009)**	**0.733 (± 0.013)**	0.72 (± 0.024)	0.594 (± 0.012)	0.438 (± 0.008)	0.621 (± 0.011)
NT-Multispecies (2.5B)	**0.743 (± 0.009)**	**0.747 (± 0.014)**	**0.774 (± 0.036)**	0.664 (± 0.007)	0.491 (± 0.01)	0.687 (± 0.004)

Supplementary Table 6: Downstream performance per task for all probing models. Every NT model was evaluated through probing and all "BPNet" models were fully trained CNN networks. Performance per task was calculated as the median of the 10 cross-validation folds (± standard deviation).

Dataset Model	H2AFZ	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me2
Raw Probe	0.302 (± 0.003)	0.229 (± 0.006)	0.379 (± 0.004)	0.36 (± 0.003)	0.287 (± 0.006)	0.36 (± 0.005)
BPNet (original)	0.473 (± 0.009)	0.296 (± 0.046)	0.543 (± 0.009)	0.548 (± 0.009)	0.436 (± 0.008)	0.427 (± 0.036)
BPNet (large)	0.487 (± 0.014)	0.214 (± 0.037)	0.551 (± 0.009)	0.57 (± 0.009)	0.459 (± 0.012)	0.427 (± 0.025)
BPNet (X-large)	0.475 (± 0.011)	0.16 (± 0.042)	0.542 (± 0.009)	0.562 (± 0.011)	0.457 (± 0.013)	0.341 (± 0.038)
DNABERT-2	0.49 (± 0.013)	0.491 (± 0.01)	**0.599 (± 0.01)**	**0.637 (± 0.007)**	**0.49 (± 0.008)**	0.558 (± 0.013)
Enformer	**0.522 (± 0.019)**	**0.52 (± 0.015)**	0.552 (± 0.007)	0.567 (± 0.017)	**0.504 (± 0.021)**	**0.626 (± 0.015)**
HyenaDNA-1KB	0.455 (± 0.015)	0.423 (± 0.017)	0.541 (± 0.018)	0.543 (± 0.01)	0.43 (± 0.014)	0.521 (± 0.024)
HyenaDNA-32KB	0.467 (± 0.012)	0.421 (± 0.01)	0.55 (± 0.009)	0.553 (± 0.011)	0.423 (± 0.016)	0.515 (± 0.018)
NT 500M Randomly Initialized	0.281 (± 0.017)	0.233 (± 0.015)	0.364 (± 0.021)	0.305 (± 0.004)	0.228 (± 0.018)	0.317 (± 0.021)
NT 2.5B Randomly Initialized	0.291 (± 0.011)	0.257 (± 0.02)	0.379 (± 0.009)	0.324 (± 0.014)	0.271 (± 0.02)	0.33 (± 0.014)
NT-HumanRef (500M)	0.465 (± 0.011)	0.457 (± 0.01)	0.589 (± 0.009)	0.594 (± 0.004)	0.468 (± 0.007)	0.527 (± 0.011)
NT-1000G (500M)	0.464 (± 0.012)	0.458 (± 0.012)	0.591 (± 0.007)	0.581 (± 0.009)	0.466 (± 0.006)	0.528 (± 0.011)
NT-1000G (2.5B)	0.478 (± 0.012)	**0.486 (± 0.023)**	**0.603 (± 0.009)**	**0.632 (± 0.008)**	**0.491 (± 0.015)**	0.569 (± 0.014)
NT-Multispecies (2.5B)	**0.503 (± 0.01)**	0.481 (± 0.02)	**0.593 (± 0.016)**	**0.635 (± 0.016)**	**0.481 (± 0.012)**	0.552 (± 0.022)
NT-Multispecies-v2 (50M)	**0.501 (± 0.012)**	0.483 (± 0.012)	0.595 (± 0.004)	0.607 (± 0.006)	**0.478 (± 0.012)**	0.549 (± 0.014)
NT-Multispecies-v2 (100M)	**0.492 (± 0.012)**	0.487 (± 0.016)	**0.595 (± 0.013)**	0.617 (± 0.006)	**0.485 (± 0.011)**	0.551 (± 0.01)
NT-Multispecies-v2 (250M)	**0.513 (± 0.017)**	**0.497 (± 0.014)**	**0.6 (± 0.009)**	**0.636 (± 0.012)**	**0.491 (± 0.006)**	0.57 (± 0.009)
NT-Multispecies-v2 (500M)	**0.493 (± 0.01)**	0.492 (± 0.012)	**0.613 (± 0.01)**	**0.642 (± 0.015)**	**0.497 (± 0.009)**	0.564 (± 0.012)

Dataset Model	H3K4me3	H3K9ac	H3K9me3	H4K20me1	Enhancer	Enhancer types
Raw Probe	0.432 (± 0.008)	0.322 (± 0.013)	0.105 (± 0.01)	0.446 (± 0.006)	0.251 (± 0.003)	0.233 (± 0.004)
BPNet (original)	0.445 (± 0.047)	0.336 (± 0.034)	0.298 (± 0.03)	0.531 (± 0.025)	0.488 (± 0.009)	0.449 (± 0.006)
BPNet (large)	0.445 (± 0.049)	0.298 (± 0.033)	0.234 (± 0.037)	0.525 (± 0.038)	0.492 (± 0.008)	0.454 (± 0.008)
BPNet (X-large)	0.338 (± 0.076)	0.235 (± 0.06)	0.216 (± 0.085)	0.411 (± 0.054)	0.501 (± 0.013)	0.455 (± 0.012)
DNABERT-2	**0.646 (± 0.008)**	**0.564 (± 0.013)**	0.443 (± 0.025)	**0.655 (± 0.011)**	0.517 (± 0.011)	0.476 (± 0.009)
Enformer	**0.635 (± 0.019)**	**0.593 (± 0.02)**	0.453 (± 0.016)	0.606 (± 0.016)	**0.614 (± 0.01)**	**0.573 (± 0.013)**
HyenaDNA-1KB	0.596 (± 0.015)	0.484 (± 0.022)	0.375 (± 0.026)	0.58 (± 0.009)	0.475 (± 0.006)	0.441 (± 0.01)
HyenaDNA-32KB	0.603 (± 0.02)	0.487 (± 0.025)	0.419 (± 0.03)	0.59 (± 0.007)	0.476 (± 0.021)	0.445 (± 0.009)
NT 500M Randomly Initialized	0.452 (± 0.031)	0.329 (± 0.035)	0.088 (± 0.015)	0.431 (± 0.012)	0.251 (± 0.012)	0.215 (± 0.022)
NT 2.5B Randomly Initialized	0.463 (± 0.023)	0.299 (± 0.021)	0.097 (± 0.035)	0.436 (± 0.01)	0.276 (± 0.012)	0.225 (± 0.021)
NT-HumanRef (500M)	0.622 (± 0.013)	0.524 (± 0.013)	0.433 (± 0.009)	0.634 (± 0.013)	0.515 (± 0.019)	0.477 (± 0.014)
NT-1000G (500M)	0.609 (± 0.011)	0.515 (± 0.018)	0.415 (± 0.019)	0.634 (± 0.01)	0.505 (± 0.009)	0.459 (± 0.011)
NT-1000G (2.5B)	0.615 (± 0.017)	0.529 (± 0.012)	**0.483 (± 0.013)**	**0.659 (± 0.008)**	0.504 (± 0.009)	0.469 (± 0.005)
NT-Multispecies (2.5B)	0.618 (± 0.015)	0.527 (± 0.017)	0.447 (± 0.018)	**0.65 (± 0.014)**	0.527 (± 0.012)	0.484 (± 0.012)
NT-Multispecies-v2 (50M)	**0.627 (± 0.013)**	0.538 (± 0.011)	0.43 (± 0.018)	0.637 (± 0.008)	0.511 (± 0.006)	0.459 (± 0.006)
NT-Multispecies-v2 (100M)	**0.633 (± 0.015)**	0.538 (± 0.015)	0.445 (± 0.017)	**0.648 (± 0.008)**	0.507 (± 0.009)	0.465 (± 0.009)
NT-Multispecies-v2 (250M)	**0.64 (± 0.009)**	**0.565 (± 0.021)**	**0.467 (± 0.016)**	**0.652 (± 0.006)**	0.525 (± 0.01)	0.492 (± 0.01)
NT-Multispecies-v2 (500M)	**0.636 (± 0.016)**	0.539 (± 0.013)	**0.475 (± 0.014)**	**0.658 (± 0.008)**	0.524 (± 0.013)	0.498 (± 0.009)

Dataset Model	Promoter all	Promoter non-TATA	Promoter TATA	Splice acceptor	Splice site all	Splice donor
Raw Probe	0.598 (± 0.002)	0.619 (± 0.004)	0.606 (± 0.007)	0.136 (± 0.04)	0.234 (± 0.003)	0.395 (± 0.025)
BPNet (original)	0.696 (± 0.026)	0.717 (± 0.023)	0.848 (± 0.042)	0.859 (± 0.038)	0.793 (± 0.072)	0.92 (± 0.014)
BPNet (large)	0.672 (± 0.023)	0.672 (± 0.043)	0.826 (± 0.017)	**0.925 (± 0.031)**	0.865 (± 0.026)	0.93 (± 0.021)
BPNet (X-large)	0.579 (± 0.087)	0.608 (± 0.061)	0.607 (± 0.064)	0.744 (± 0.01)	0.745 (± 0.016)	0.706 (± 0.009)
DNABERT-2	0.754 (± 0.009)	**0.769 (± 0.009)**	0.784 (± 0.036)	0.837 (± 0.006)	0.855 (± 0.005)	0.861 (± 0.004)
Enformer	0.745 (± 0.012)	0.763 (± 0.012)	0.793 (± 0.026)	0.749 (± 0.007)	0.739 (± 0.011)	0.78 (± 0.007)
HyenaDNA-1KB	0.693 (± 0.016)	0.723 (± 0.013)	0.648 (± 0.044)	0.815 (± 0.049)	0.854 (± 0.053)	0.943 (± 0.024)
HyenaDNA-32KB	0.698 (± 0.011)	0.729 (± 0.009)	0.666 (± 0.041)	0.808 (± 0.009)	0.907 (± 0.018)	0.915 (± 0.047)
NT 500M Randomly Initialized	0.575 (± 0.01)	0.581 (± 0.025)	0.39 (± 0.044)	0.207 (± 0.031)	0.363 (± 0.038)	0.596 (± 0.057)
NT 2.5B Randomly Initialized	0.589 (± 0.014)	0.597 (± 0.022)	0.465 (± 0.04)	0.258 (± 0.049)	0.421 (± 0.024)	0.641 (± 0.045)
NT-HumanRef (500M)	0.734 (± 0.013)	0.738 (± 0.008)	0.831 (± 0.022)	0.941 (± 0.004)	0.939 (± 0.003)	0.952 (± 0.003)
NT-1000G (500M)	0.727 (± 0.004)	0.743 (± 0.012)	0.855 (± 0.041)	0.933 (± 0.007)	0.939 (± 0.004)	0.952 (± 0.004)
NT-1000G (2.5B)	0.708 (± 0.008)	0.758 (± 0.007)	0.802 (± 0.03)	**0.952 (± 0.004)**	0.956 (± 0.004)	0.963 (± 0.001)
NT-Multispecies (2.5B)	**0.761 (± 0.009)**	**0.773 (± 0.01)**	**0.944 (± 0.016)**	**0.958 (± 0.003)**	**0.964 (± 0.003)**	**0.97 (± 0.002)**
NT-Multispecies-v2 (50M)	0.731 (± 0.01)	0.756 (± 0.01)	0.747 (± 0.043)	0.922 (± 0.006)	0.928 (± 0.005)	0.915 (± 0.006)
NT-Multispecies-v2 (100M)	0.753 (± 0.005)	**0.766 (± 0.014)**	0.826 (± 0.019)	0.947 (± 0.003)	**0.96 (± 0.005)**	0.947 (± 0.008)
NT-Multispecies-v2 (250M)	**0.774 (± 0.013)**	**0.785 (± 0.01)**	0.87 (± 0.019)	**0.95 (± 0.008)**	**0.965 (± 0.003)**	**0.967 (± 0.004)**
NT-Multispecies-v2 (500M)	**0.778 (± 0.012)**	**0.774 (± 0.015)**	0.878 (± 0.021)	0.944 (± 0.005)	**0.967 (± 0.004)**	**0.96 (± 0.01)**

Supplementary Table 7: Downstream performance per task for all finetuned models. Every model was fine-tuned with PEFT, except for "BPNet" models that are fully trained CNN networks and "Raw Probe" that used probing strategy. Performance per task was calculated as the median of the 10 cross-validation folds (± standard deviation).

	Model	Layer probed	Number of hidden layers	Hidden layer dimension	Batch size	Learning rate
Enhancer	2.5B 1000G	17	1	256	500	1e-4
H4K20me1	2.5B 1000G	17	1	256	500	2e-4
H2AFZ	2.5B 1000G	17	1	256	500	1e-4
H3K4me1	2.5B 1000G	14	2	512	500	1e-4
H3K9ac	2.5B MS	10	1	256	500	1e-4
Promoter all	2.5B MS	10	1	512	500	3e-4
Promoter non-TATA	2.5B MS	10	1	256	500	1e-4
H3K36me3	2.5B 1000G	17	1	256	500	1e-4
Enhancer types	2.5B MS	6	1	256	500	1e-4
H3K27me3	2.5B 1000G	17	1	256	500	1e-4
Splice acceptor	2.5B MS	10	1	256	500	1e-4
H3K4me3	2.5B MS	17	1	256	500	1e-4
Promoter TATA	2.5B MS	10	2	512	500	5e-4
Splice donor	2.5B MS	10	1	512	500	1e-4
H3K4me2	2.5B 1000G	17	1	256	500	1e-4
H3K9me3	2.5B MS	10	1	256	500	1e-4
H3K27ac	2.5B 1000G	17	1	256	500	1e-4
Splice site all	2.5B MS	14	1	512	652	1e-4

Supplementary Table 8: Hyperparameters used to train the probing model with the best cross-validation performance for each of the downstream task, along with the model and the layer used to compute the corresponding embeddings. The best probe method for the tasks **Enhancer types** and **Splice donor** was the logistic regression with default hyperparameters.

Model name	Model type	Approach	Pre-training	Average performance
NT 500M Randomly Initialized	Control	Training from scratch	NA	0.355
Raw Probe	Raw Probe	Probing	Self-supervised	0.391
NT 2.5B Randomly Initialized	Control	Training from scratch	NA	0.399
DNABERT-1	DNABERT-1	Finetuning	Self-supervised	0.511
NT-HumanRef (500M)	NT-v1	Probing	Self-supervised	0.534
BPNet (X-large)	BPNet	control	NA	0.566
NT-1000G (500M)	NT-v1	Probing	Self-supervised	0.583
NT-1000G (2.5B)	NT-v1	Probing	Self-supervised	0.613
NT-Multispecies (2.5B)	NT-v1	Probing	Self-supervised	0.644
BPNet (original)	BPNet	control	NA	0.665
HyenaDNA-1KB	HyenaDNA	Finetuning	Self-supervised	0.668
BPNet (large)	BPNet	control	NA	0.683
Enformer	Enformer	Finetuning	Supervised	0.684
HyenaDNA-32KB	HyenaDNA	Finetuning	Self-supervised	0.691
DNABERT-2	DNABERT-2	Finetuning	Self-supervised	0.723
NT-1000G (500M)	NT-v1	Finetuning	Self-supervised	0.729
NT-Multispecies-v2 (50M)	NT-v2	Finetuning	Self-supervised	0.732
NT-HumanRef (500M)	NT-v1	Finetuning	Self-supervised	0.733
NT-1000G (2.5B)	NT-v1	Finetuning	Self-supervised	0.738
NT-Multispecies-v2 (100M)	NT-v2	Finetuning	Self-supervised	0.748
NT-Multispecies (2.5B)	NT-v1	Finetuning	Self-supervised	0.755
NT-Multispecies-v2 (500M)	NT-v2	Finetuning	Self-supervised	0.762
NT-Multispecies-v2 (250M)	NT-v2	Finetuning	Self-supervised	**0.769**

Supplementary Table 9: Normalized summed downstream performance of all models, sorted by performance. These values are obtained by averaging the median MCC score obtained by each model on the three categories of downstream tasks: Chromatin Profiles, Regulatory elements and Splicing. Information about model type and probing/fine-tuning approach is also provided.

	Percentage of sequences containing this element	Mean Percentage of Sequence Length belonging to this element
Enhancer	10.18%	11.09%
Open chromatin	11.32%	6.16%
3' UTR	11.37%	4.30%
TF binding	11.28%	6.38%
Intron	10.45%	11.08%
Exon	11.87%	3.18%
5'UTR	11.84%	1.99%
Promoter	10.95%	9.78%
CTCF Binding Site	10.69%	7.36%

Supplementary Table 10: Proportions of the genomic elements in the dataset used for attention maps and embedding space visualization experiments.

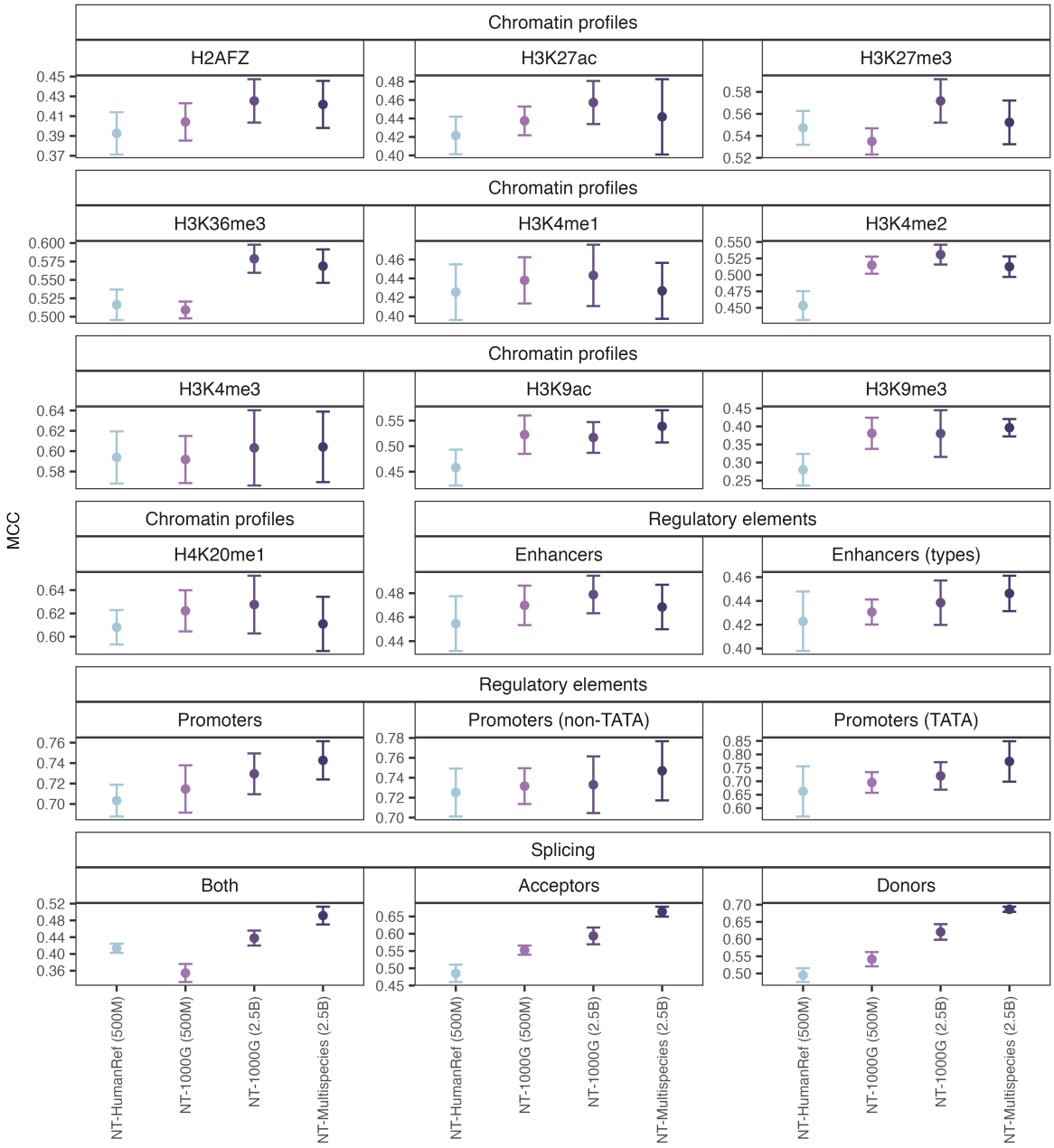
	CDS	CTCF	Enhancer	Exon	5'UTR	Intron	Open chromatin	Promoter	TF	3'UTR	No. heads
Human ref (500M)	40	26	27	33	37	37	27	33	26	37	480
1000G (500M)	52	27	24	48	41	40	26	55	27	49	480
Multispecies (2.5B)	89	35	36	72	42	117	44	50	74	51	640
1000G (2.5B)	38	32	28	40	33	61	40	55	45	54	640

Supplementary Table 11: Number of significant attention heads capturing gene features and regulatory elements. Note that the number of total heads varies between the 500M and 2.5B models.

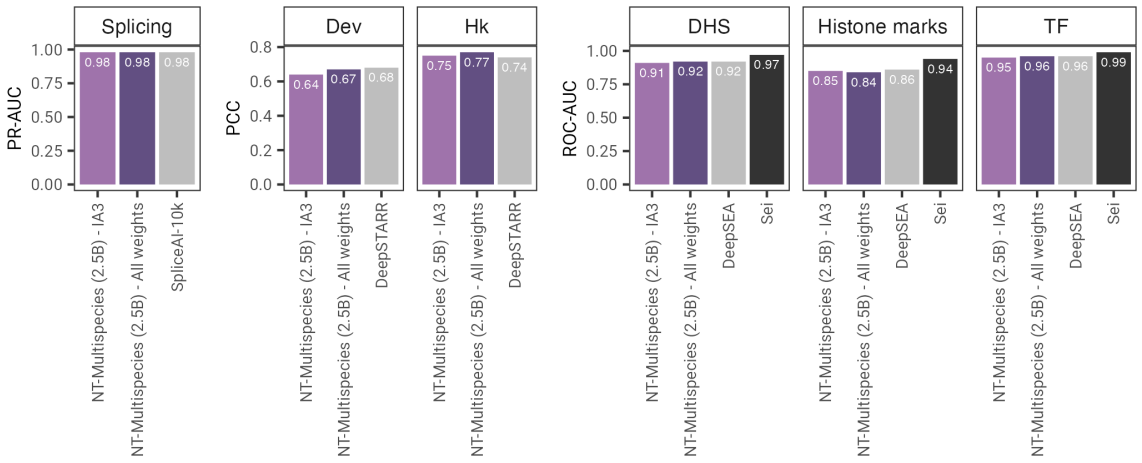
	CDS	CTCF	Enhancer	Exon	5'UTR	Intron	Open chromatin	Promoter	TF	3'UTR	No. heads
Human ref (500M)	40	26	27	33	37	37	27	33	26	37	480
1000G (500M)	52	27	24	48	41	40	26	55	27	49	480
Multispecies (2.5B)	89	35	36	72	42	117	44	50	74	51	640
1000G (2.5B)	38	32	28	40	33	61	40	55	45	54	640

Supplementary Table 12: Number of significant attention heads capturing gene features and regulatory elements. Note that the number of total heads varies between the 500M and 2.5B models.

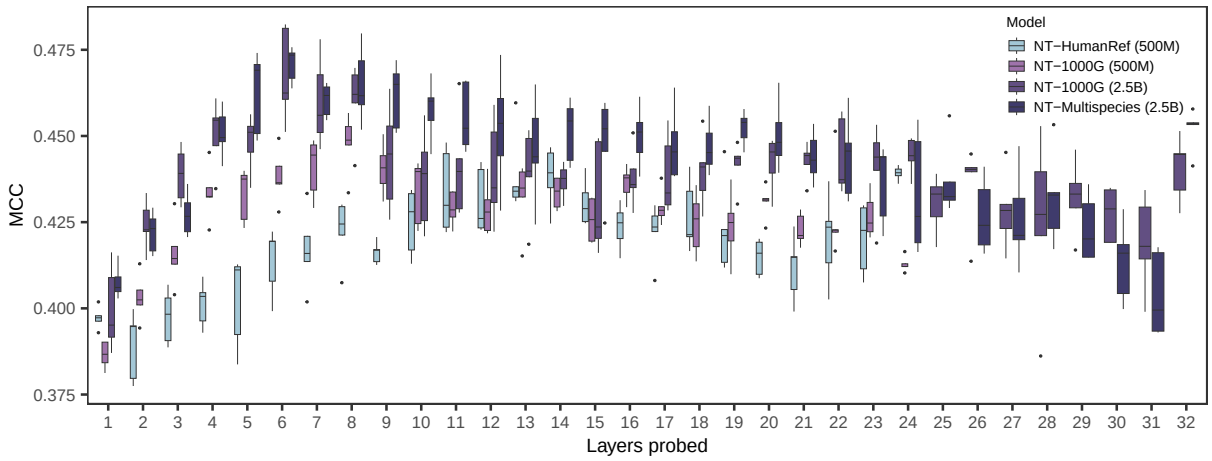
C Supplementary Figures



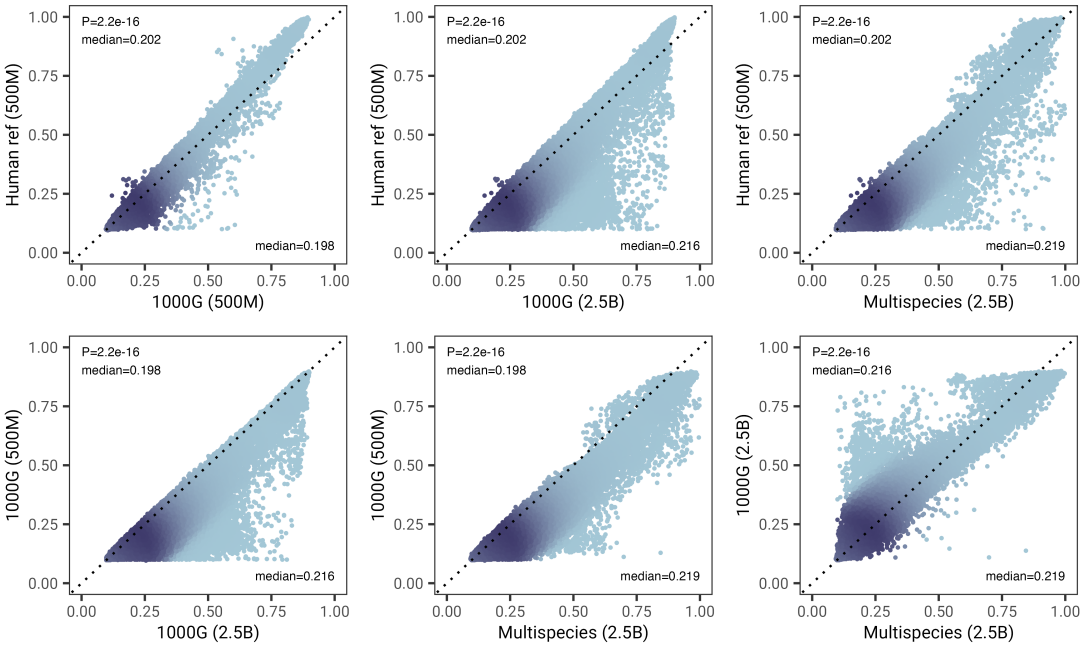
Supplementary Figure 1: Probing performance results on test sets across downstream tasks for each Nucleotide Transformer model. The performances of the best layers and best downstream models are shown. Data are presented as mean MCC values +/- 2 standard deviations from the 10-fold cross-validation procedure (n=10 for each point).



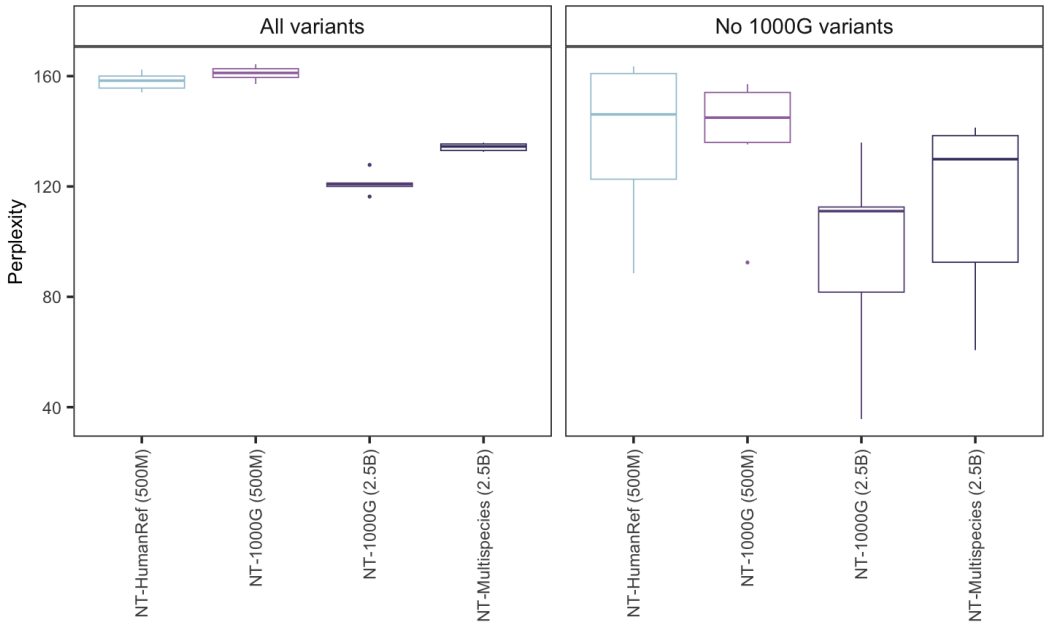
Supplementary Figure 2: Parameter efficient fine-tuning compared with full-model fine-tuning. The performances across different tasks are shown and compared with respective baselines.



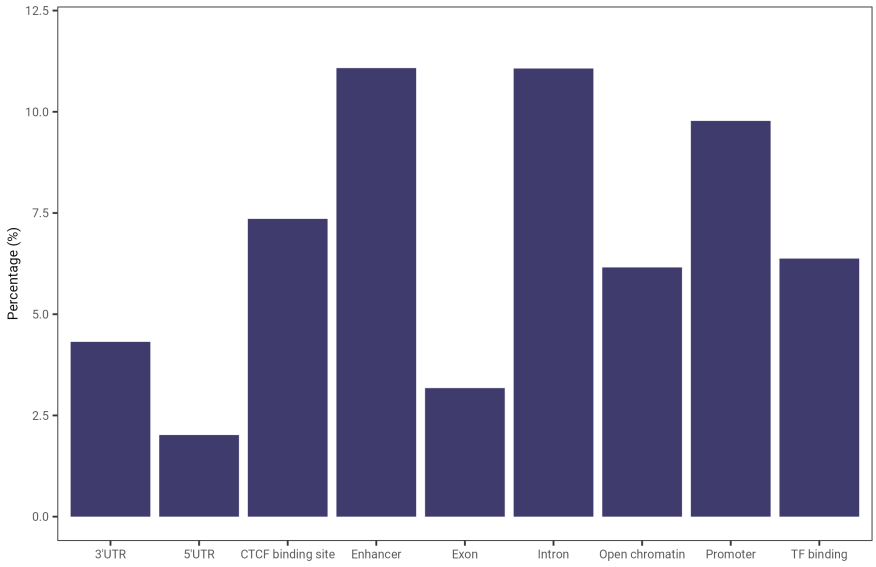
Supplementary Figure 3: Probing performance across layers for the enhancer (types) prediction task. Boxplots show the MCC values across layers based on 10 fold cross-validation experiments (n=10 for each boxplot). The boxplots mark the median, upper and lower quartiles, and 1.5× interquartile range (whiskers).



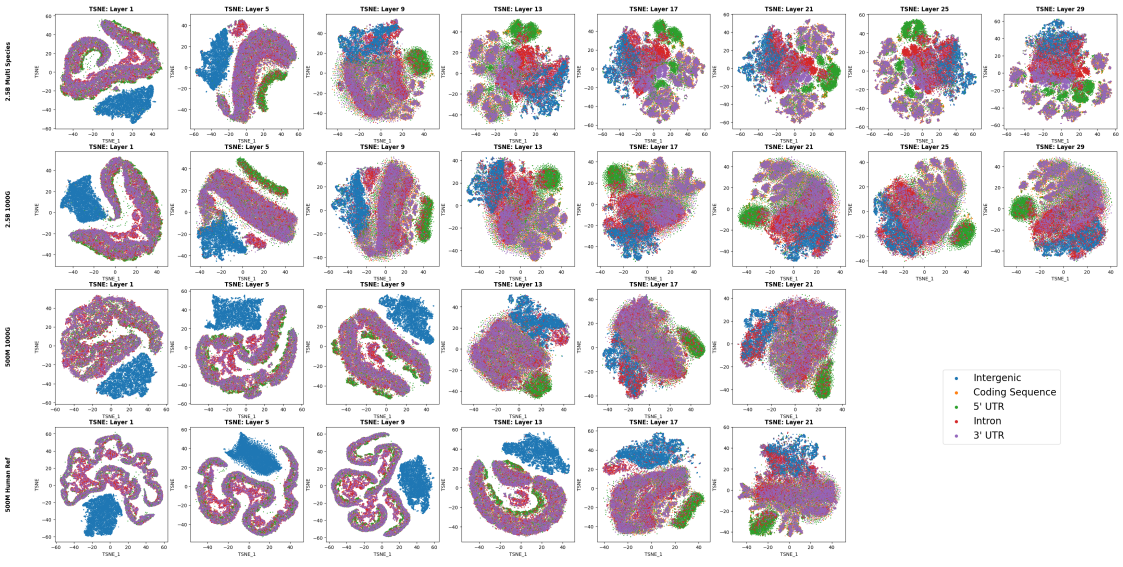
Supplementary Figure 4: Pairwise comparison of token reconstruction accuracy across Nucleotide Transformer models. P-values refer to a two-sided Wilcoxon signed rank test between models. Median values for the two models compared are shown.



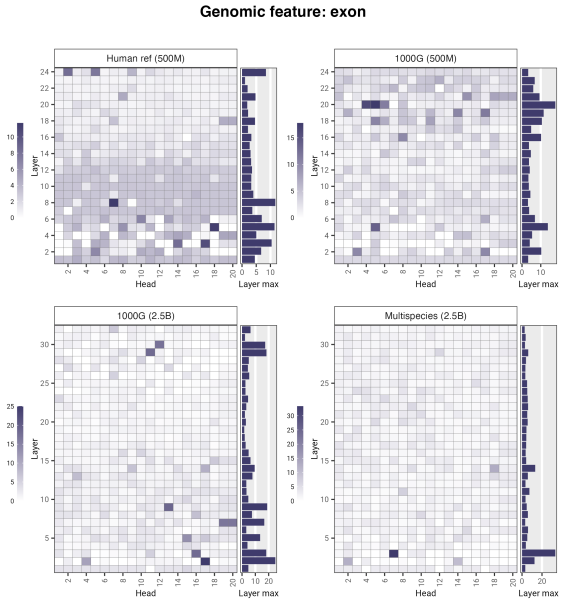
Supplementary Figure 5: Reconstruction perplexity per model for variants present in 6 meta-populations from the Human Genome Diversity Project. Perplexity values are based on 6 meta-populations from the Human Genome Diversity Project (n=6 per boxplot). Results are shown for all variants or for variants that are not present in the 1000G samples used for model pre-training. The boxplots mark the median, upper and lower quartiles, and 1.5× interquartile range (whiskers).



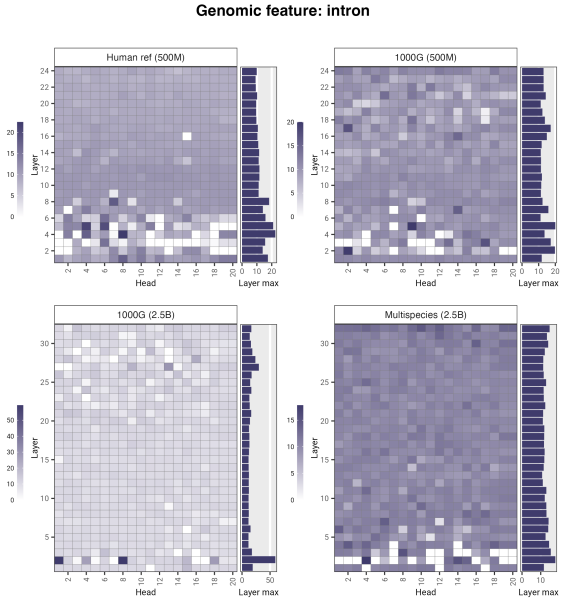
Supplementary Figure 6: Genomic element length's percentage of input sequence (6kbp). This corresponds to the dataset used for attention maps and embedding spaces visualization experiments. Details in Table 10.



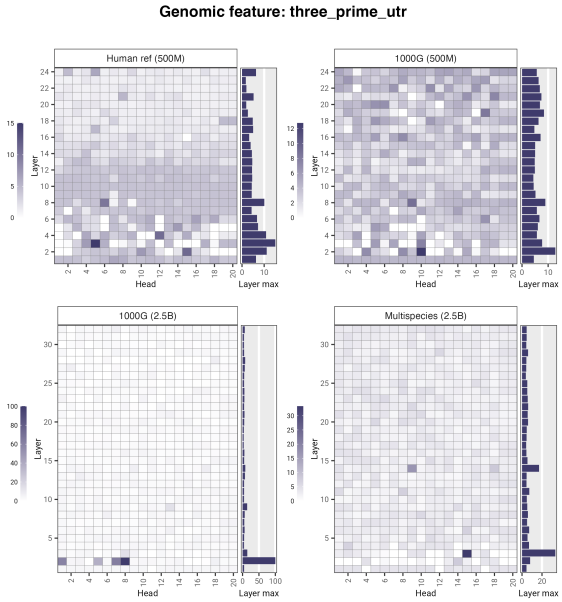
Supplementary Figure 7: t-SNE projections of embeddings of 5 genomic elements across transformer models and layers.



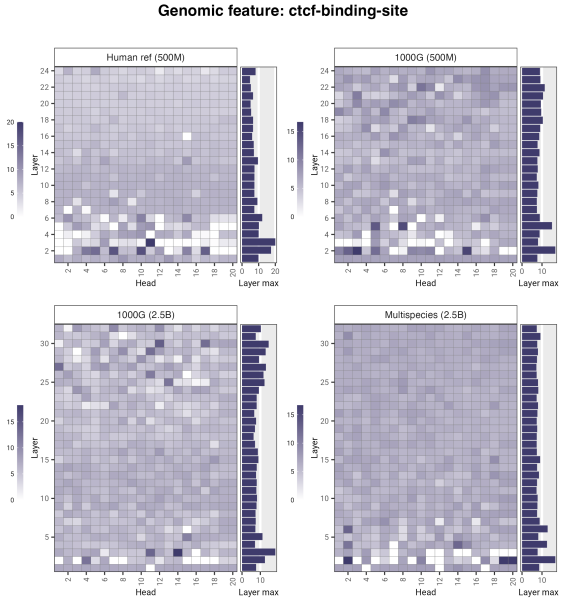
Supplementary Figure 8: Nucleotide Transformer models' attention percentages per head computed for exons.



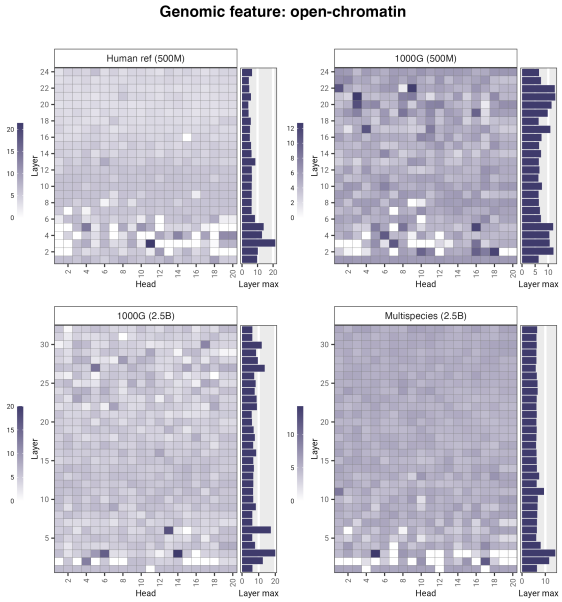
Supplementary Figure 9: Nucleotide Transformer models' attention percentages per head computed for introns.



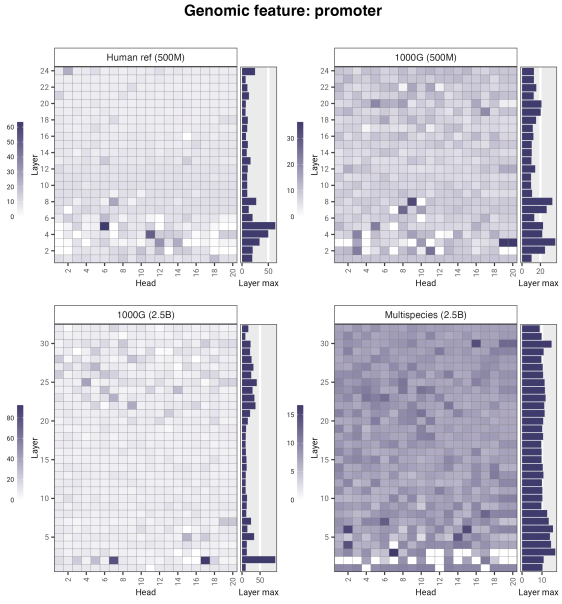
Supplementary Figure 10: Nucleotide Transformer models' attention percentages per head computed for 3' UTR regions.



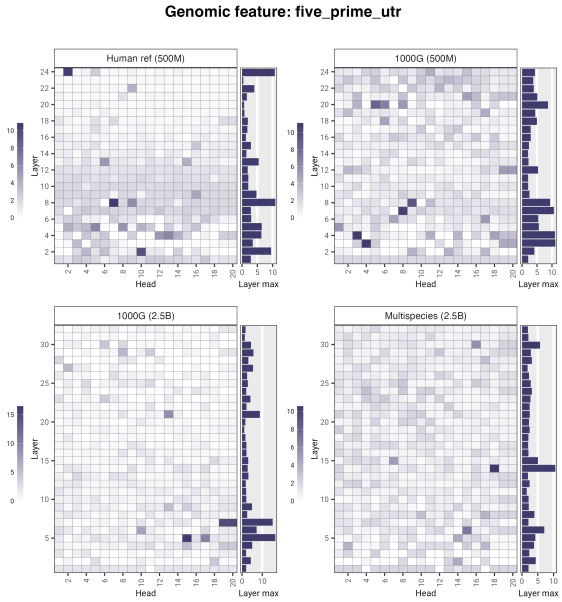
Supplementary Figure 11: Nucleotide Transformer models' attention percentages per head computed for CTCF binding sites.



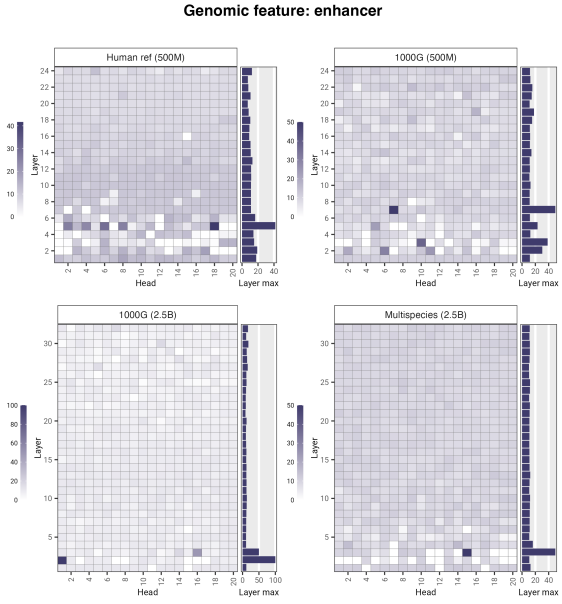
Supplementary Figure 12: Nucleotide Transformer models' attention percentages per head computed for open chromatin.



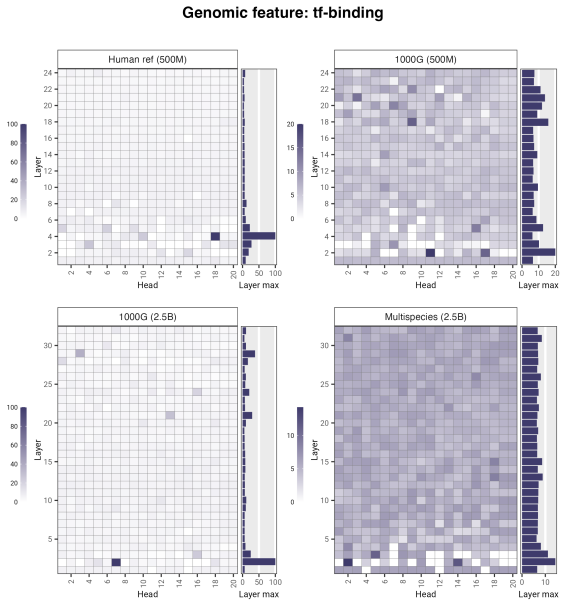
Supplementary Figure 13: Nucleotide Transformer models' attention percentages per head computed for promoters.



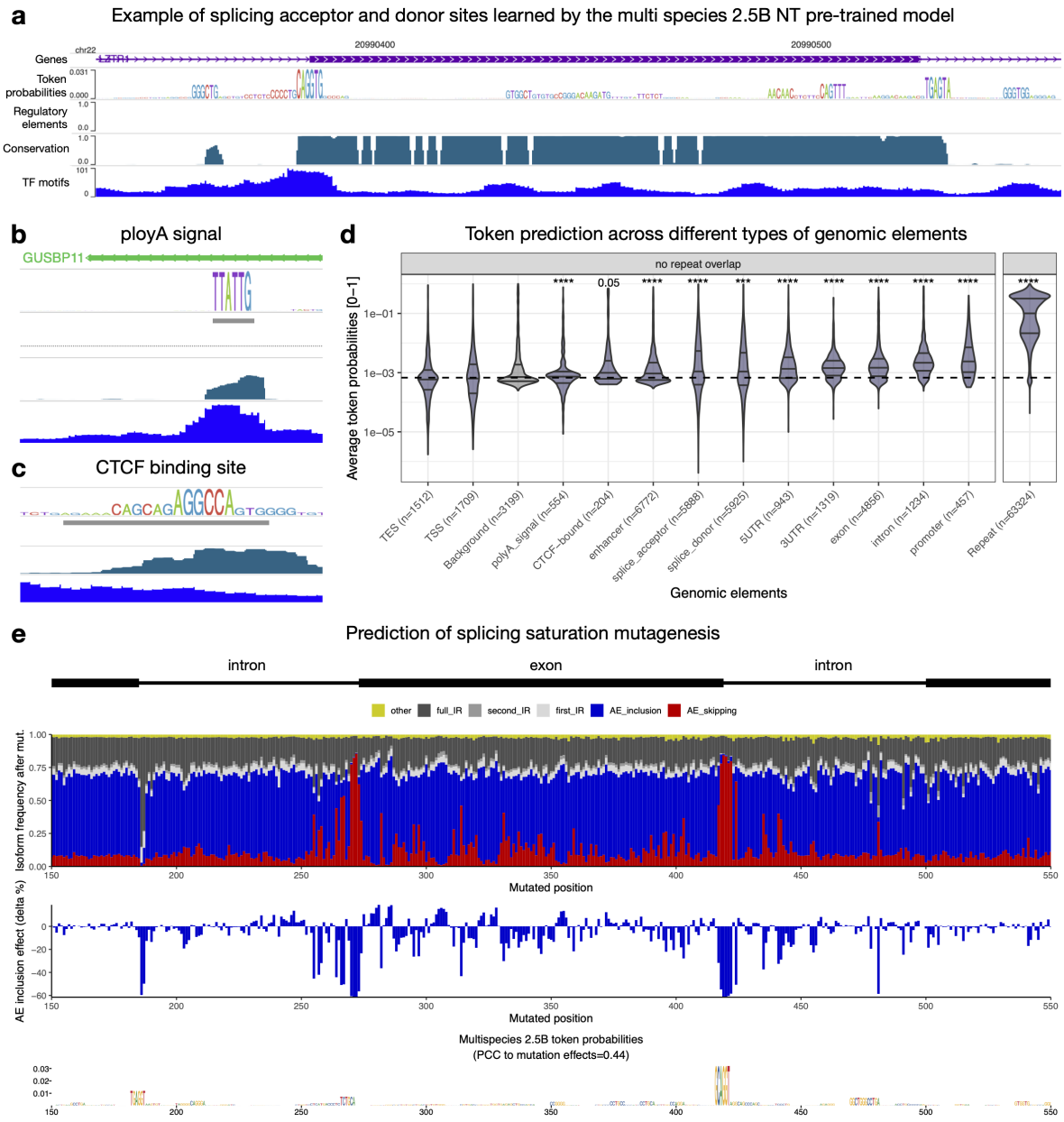
Supplementary Figure 14: Nucleotide Transformer models' attention percentages per head computed for 5' UTR regions.



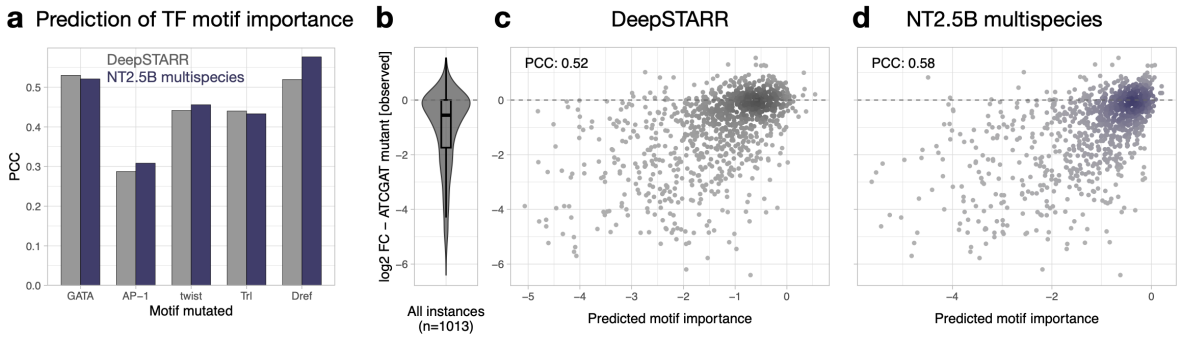
Supplementary Figure 15: Nucleotide Transformer models' attention percentages per head computed for enhancers.



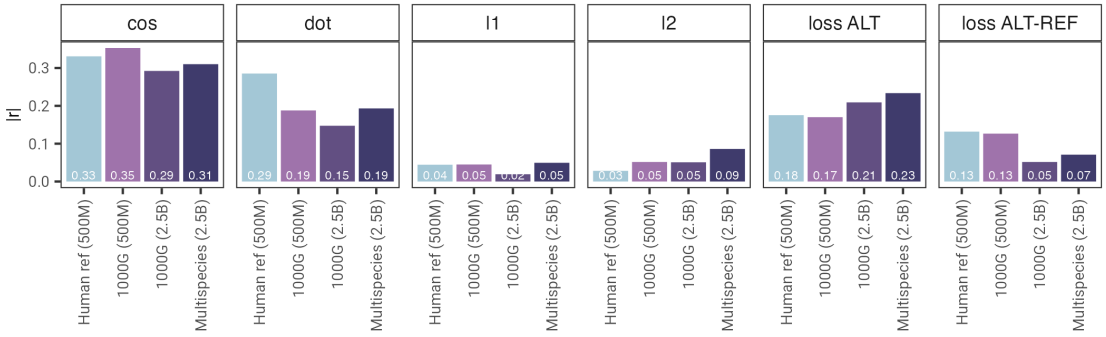
Supplementary Figure 16: Nucleotide Transformer models' attention percentages per head computed for transcription factor binding sites.



Supplementary Figure 17: Interpretation of Nucleotide Transformer multispecies 2.5B pre-trained model at higher resolution. **a-c)** Genome browser screenshot depicting token probabilities for **(a)** an exon region of gene LZTR1 (chr22:20990296-20990630), **(b)** a polyA signal from GUSBP11 3'UTR (chr22:23638476-23638540), and **(c)** a CTCF binding site (chr22:22254952-22254981). ENCODE regulatory elements, conservation, and Jaspar TF motifs are also shown. **d)** Distribution of average token probability scores across different types of genomic elements that do not overlap repeat elements (left) or across repeat elements (right). Violins are shown with horizontal lines marking the median and lower and upper quartile. ****P ¡ 0.0001, ***P ¡ 0.001, **P ¡ 0.01, *P ¡ 0.05, (two-sided Wilcoxon rank-sum test). **e)** Top: experimentally measured impact of mutating every nucleotide position in the different splicing isoforms of exon 11 of gene MST1R. Middle: effect in Alternative Exon (AE) 11 inclusion as delta percentage. Bottom: token probability predictions by the multispecies 2.5B pre-trained model. Pearson Correlation Coefficient (PCC) between token predictions and AE11 inclusion mutation effects is shown.



Supplementary Figure 18: Prediction of transcription factor motif importance. **a)** Bar plots showing the Pearson Correlation Coefficient (PCC) between predicted (by DeepSTARR or the multispecies 2.5B full-finetuned model) and observed log2FC for mutating individual instances of each motif type. **b-d)** Distribution of experimentally measured enhancer activity log2FC after mutating 1,013 different Dref motif instances across enhancers (violin plot), compared with the log2FC predicted by **(c)** DeepSTARR or **(d)** the multispecies 2.5B model. The boxplot marks the median, upper and lower quartiles, and 1.5× interquartile range (whiskers).



Supplementary Figure 19: Performance of zero-shot based scores for SNPs annotations based on Ensembl Variant Effect Prediction (VEP). Performance is based on the Pearson correlation between the scores and severity as estimated by Ensembl.