Ge3Net: Inferring Continuous Population Structure Coordinates Along the Genome

Richa Rastogi[1,*], Helgi Hilmarsson[3], Arvind S. Kumar[3], Carlos D. Bustamante[2], Daniel Mas Montserrat[2], and Alexander G. Ioannidis[2,3]

[1]Cornell University, Department of Computer Science, New York, 10044, USA
[2]Stanford Medical School, Department of Biomedical Data Science, Stanford, 94305, USA
[3]Stanford University, Institute for Computational and Mathematical Engineering, Stanford, 94305, USA
[*]corresponding author: rr568@cornell.edu
[*]Work done while at Stanford (2020-2021)

ABSTRACT

Personalized genomic predictions are revolutionizing medical diagnosis and treatment. These predictions rely on associations between health outcomes (disease severity, drug response, cancer risk) and correlated neighboring positions along the genome. However, these local genomic correlations differ widely amongst worldwide populations, necessitating that genetic research include all human populations. For admixed populations further computational challenges arise, because individuals of diverse combined ancestries inherit genomic segments from multiple ancestral populations. To extend population-specific associations to such individuals, their multiple ancestries must be identified along their genome (local ancestry inference, LAI). Here we introduce Ge3Net, Genomic Geographic Geometric Network, the first LAI method to identify ancestral origin of *each segment* of an individual's genome as a *continuous coordinate*, rather than an ethnic category, using a transformer based framework, yielding higher resolution *local* ancestry inference, and eliminating a need for ethnic labels.

1 Background & Summary

Precision medicine aims to break free from traditional one-size-fits-all treatments by providing personalized solutions tailored to the biology of each individual. Genomic medicine plays an important role in this, by enabling the targeting of treatments and risk assessments tailored to a patient's genetics. Genome-wide association studies (GWAS) are used to identify these relationships between variation in an individual's genome and drug response, disease predisposition, or other traits. GWAS exploit correlations (linkage) between sequenced (genotyped) variants and the neighbouring sites on the DNA strand that cause a condition or trait. These correlations differ between long-separated populations due to genetic drift[1,2], meaning that association models (known as polygenic risk scores, or PRS) designed for one population fail when applied to another[3–5]. Indeed, in some cases the same variant at the same position in the DNA strand can have an opposite association with a disease, depending on the ancestry of the genome (individual) in which it is found[6,7]. Despite the imperative this phenomenon places on studying individuals from a wide range of ancestries, most GWAS to-date focus on European-descendent groups[8,9]. Thus, by excluding the majority the world's ancestries from such studies, a new divide is opening in next generation healthcare.

This divide, affecting which populations will benefit from accurate genetic association models, is not emerging through a scarcity of available participants from diverse populations, but rather because most studies focus on single ancestry cohorts to increase statistical power and to eliminate the potential for these opposing associations in different ancestries. Admixed groups (such as African-American and Latin American populations), which have multiple ancestries (European and African or Native American, respectively), are often deliberately excluded. This approach is not sustainable, as inhabitants of the developed world become increasingly admixed; indeed, over half of all newborns in California have parents from more than one racial or ethnic group (white, black, Asian/Pacific Islander, American Indian, and Hispanic)[10]. Moreover, recent work[11] argues how a race oriented interpretation might arise from continental classification[12–18].Thus, there is a need for algorithms, such as local ancestry inference-based methods, that can model ancestry as a continuous spectrum and aid in the inclusion of complex admixed populations in association studies (GWAS). This will further enable constructing genetic risk prediction models (polygenic risk scores, PRS) by providing ancestry as an additional covariate along the genome for admixed individuals.

Local-ancestry inference (LAI) methods, also known as ancestry deconvolution methods, have traditionally labeled each region of the genome with an ancestry category. These ancestry labels can be used as co-variates in association studies to de-convolve the effects of ancestry on local genomic correlation structure and on trait associations[19]. Multiple methods for classification-based local-ancestry inference have been introduced in recent years: SABER[20], HAPAA[21], HAPMIX[22], and

LAMP[23] make use of Hidden Markov Models. SVMs[24] and Random Forests (RFMix)[25] have provided higher accuracy and faster execution times. Neural network-based methods have also been explored with LAI-Net[26] providing results competitive with RFMix.

Several neural network-based techniques for processing population genetics data have been presented recently[27]. Some examples include Neural ADMIXTURE[28] for unsupervised global population clustering, generative neural networks[29,30] for genotype simulation, Diet Network[31] for population classification, PG-GAN[32] for demographic parameter inference.

However, the ancestry labels for classification based ancestry inference are selected somewhat artificially, typically assigning ethnic classifications (such as European, South Asian, and East Asian) that neglect all variability within the class and ignore the continuum of human variation that exist spatially between the discrete classes. Recent work[33], has emphasized on the need for a continuous view of ancestry and urged the field to move away from continental classification categories. For global ancestry inference, a few geographical coordinate methods that predict the origin of a single ancestry (non-admixed) individual based on their complete diploid genome exist. Examples include SPA[34], which takes a statistical approach, and Locator[35], which uses a neural network. However, to our knowledge, no method to date can infer ancestry along short segments of the genome as a continuous spectrum for any coordinate system, including geography as this setting (local ancestry inference on a continuous spectrum) has been intractable.

In this work we re-formulate the local ancestry task as one of assigning geographic coordinates (e.g. longitude and latitude) to each chromosomal segment. This task, which we refer to as spatial local-ancestry inference (Spatial LAI), has two main advantages: first it is able to resolve the ancestry differences within the same continent that regional classes ignore, and second, it avoids the use of arbitrary ethnic labels and regional divisions and can instead describe the full continuity of worldwide genetic variation. Furthermore, we explore the prediction of coordinates in euclidean spaces that properly capture the genetic variation of samples. While geographical coordinates capture most of the variation of genetic information, they fall short for some populations and are even not applicable for some species (e.g. domesticated dogs) where artificial breeding has removed the correlation between geography and genetic variability. Therefore, we explore the local prediction of coordinates constructed by using PCA or UMAP from all chromosomal sequences. Such predictions try to assign a continuous coordinate to each SNP windows such that the variability between populations is locally captured, therefore being able to provide high-resolution continuous LAI predictions for admix individuals without the need of relying on the geographic distribution. We introduce Ge3Net - Genomic Geographical Geometric Network, a neural network framework consisting of transformer encoder followed by bi-directional LSTM at its core, and designed to address Spatial LAI, providing a high-resolution ancestry prediction for each small segment of each haploid chromosome, as required for medical applications (such as GWAS and PRS).

2 Results

2.1 Ge3Net Overview

We introduce Ge3Net (Figure 1 (A)), a **Ge**ometric and **Ge**ographic **Ge**nomic Neural Network that performs continuous, coordinate-based local ancestry inference. Unlike existing local ancestry methods, which predict categorical ancestry labels along the genome, Ge3Net predicts ancestry at each position along the genome from a smooth set of coordinates that can represent the population structure ancestral to each piece of an individual's chromosomal sequence at high resolution within the full continuum of present-day (or ancient) population variation. This continuous, coordinate-based approach to local ancestry differs from the existing, discrete, label-based approaches that associate each segment of the genome instead with pre-defined, socially constructed, categories that fail to capture the spectrum and fine-grained diversity of real population structure. We showcase the flexibility of the system by performing regression of geographical coordinates and abstract coordinates within an euclidean space created with dimensional reduction techniques such as PCA or UMAP. The neural network is trained end-to-end with supervised learning with sequences from simulated admixed individuals obtained by combining single-ancestry samples from which geographical coordinates (encoded as n-vectors) are available. Alternatively, geographical coordinates can be replaced by geometric coordinates (of any dimensionality) obtained with unsupervised dimensionality reduction methods that capture genetic variation. Details on how the space is constructed are provided in 2.5.

The network takes as input haploid sequences of bi-allelic Single Nucleotide Polymorphisms (SNPs) which are the positions of the genome that are known to change between individuals. The common variant, typically referred to as reference, is encoded as a 0, and the minority variation, also named alternative, is encoded as a 1. The maternal and paternal chromosomes are treated independently. Each sequence is divided into non-overlapping windows of 1000 SNPs and a different set of multilayer perceptrons (MLPs) consisting of fully-connected layers followed by ReLU activations are applied to each window. Note that the input data is not translation invariant (the same sequence can indicate different ancestral origins at different genomic positions), therefore window-based approaches are commonly used[25]. The output of each MLP is fed into the next layers consisting of a transformer layer and a biLSTM layer. Specifically, a positional encoding for each window segments are applied before passing through a Transformer Encoder layer[36]. The self-attention module in the transformer layer enables each window segment to attend to every other window segment. The outputs of the transformer encoder layers within each window are

processed with a bi-directional LSTM. Because neighbouring regions of the chromosome are inherited together with only occasional breaks from recombination, contiguous windows are likely to have the same ancestral origin. The biLSTM allows us to capture this interaction between neighboring windows while still retaining global information about the sequence. Previous window-based classification methods have used conditional random fields[25] and convolutional layers[26] to combine predictions of multiple windows. Our method differs from such approaches as we don't use predictions, but rather hidden representations (the output of attention layers), while the biLSTM provides more flexibility than CRFs and captures longer relationships than shallow convolutional layers. A secondary auxiliary branch is included to obtain more accurate results and a more stable training. It has been shown in deep learning literature[37] that adding a small auxiliary network can help the main network by acting as a regularizer and aid in the flow of gradients during training.

2.2 Learnt Representations

In order to better understand how the network internally represents genetic ancestry and how the genotype input changes as it is passed through the different modules of Ge3Net, we perform dimensionality reduction with PCA and plot the first three PC components corresponding to each module in Figure 1(A) - (E). The PCA of the windowed input (Figure 1 (B)) shows that all the ancestries are overlapping and cannot be easily separated. This is expected as the data is not translation invariant (the same sequence in different windows will indicate different ancestries). Figure 1 (C) shows the PCA of the outputs from the first window-specific Fully Connected (FC) Layer. By applying a different linear transformation within each windows, the network is capable of learning a representation that can accurately separate ancestries on a continental level. Figure 1 (D) shows the PCA after the transformer encoder layer which provides a more fine-grained sub-continental separability and a spherical geographical structure starts to emerge. Figure 1 (E) shows the PCA after the biLSTM layer where the clusters start to become more compact and subpopulations can be separated. Figure 1 (F) shows the N-vector output of the test set samples with Ge3Net. We can understand the first linear layer at each window as a type of Factor Analysis, where the visible inputs (window of SNP sequences) are mapped to hidden factors (a shared ancestry representation) through a linear mapping. We can observe that deeper layers provide more compact and discriminative representations. The output of the transformer encoder has a spherical shape and starts to exhibit a global structure, however it is critical to note that the geographical structure is diffused, meaning the difference between ancestries as they transition from one to another is not distinct. We observe that the output of the biLSTM layers is able to discriminate and find clusters within the continental classes. Finally in Figure 1 (E), we note that the n-vectors recover the geographical structure mirroring the globe. These embeddings shows that from early layers the network can separate populations within continental groups (each with a different color). The learned embeddings are geographically consistent. That is, ancestries close in the embedding space are also close geographically. For example, the Finnish samples are located above (north) of the European (EUR) cluster, while the Spanish, Tuscan samples below (south). Similarly the Mozabite samples are located at the northern part of African continent between Southern Europe and African populations.

2.3 Ancestry Inference for Simulated Admixed

In order to train and evaluate the proposed method, genomic sequences from admixed individuals are simulated by combining genotypes from single-ancestry individuals. Specifically, admixed haplotypes were simulated by following the procedure in[38,39]. The number of changepoints are sampled from a Poission distribution with a rate given by recombination map, measured in centi-morgan (cMs) which specifies the cumulative probability of having a switch at each physical position per generation. Each single-ancestry individual (also referred to as founder) is assigned a coordinate which consists of either a geographical coordinate (n-vector) indicating where the sample was sequenced, or an abstract coordinate obtained through a dimensionality reduction technique (PCA, UMAP, etc) applied to all the single-ancestry samples. After the simulation process, each SNP from an admixed sequence is assigned the coordinate from its original ancestor.

We plot the inferred and ground truth ancestries of the simulated admixed individuals in Figure 2. For visualization purposes a categorical ancestry label is assigned to each predicted coordinate by performing a nearest neighbour mapping to the nearest training location with an L2 distance. A different color is assigned to each ground truth subpopulation label. Figure 2 (a) shows the results for a simulated admixed individual from founders belonging to Karitiana, Japanese and Mende population groups. Ge3Net is able to infer the window specific geographical (and geometrical) coordinates capturing the ancestry information, as well as ancestry changepoint (switchpoints). As the uncertainty of the predictions can be high at changepoints, such regions can be flagged or removed and allow the user to ignore them in downstream applications or analysis. The methodology for changepoints is explained in detail in 2.9. Similarly, Figure 2 (b) and (c) show Ge3net predictions in PCA space for humans and UMAP space for Canids respectively.

2.4 Ancient Genomes enable Spatio-Temporal Regression

By training a network from sequences of individuals that lived in a specific time frame and performing inference on sequences from different epochs, the genetic composition through time can be characterized. To illustrate such spatio-temporal regression

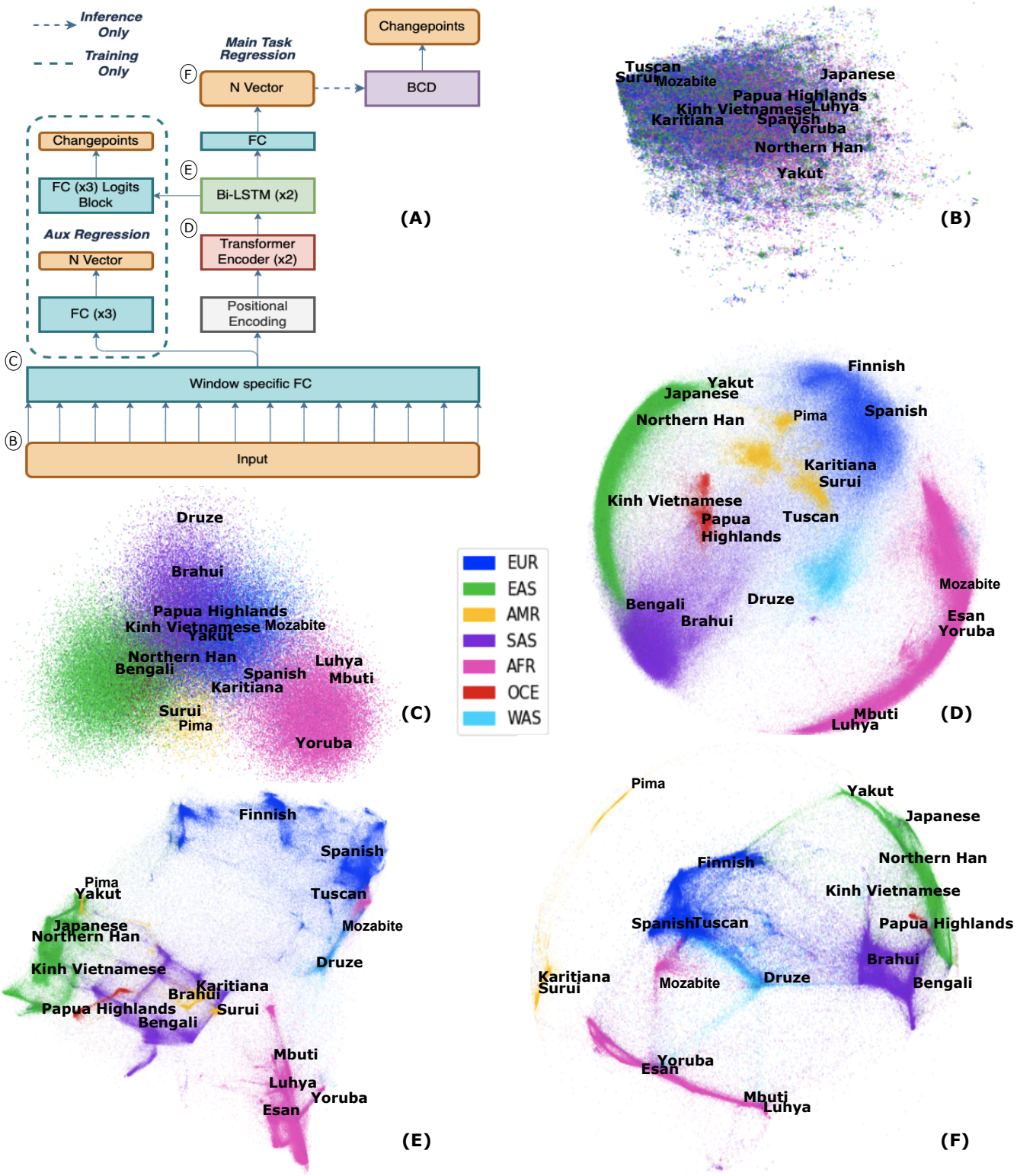

Figure 1. Learnt Representations and Architecture for Ge3Net **(A)** Ge3Net Architecture **(B)** PCA of raw admixed data **(C)** PCA after Window specific Linear layer **(D)** PCA after Transformer Encoder **(E)** PCA after BiLSTM layer **(F)** Ge3Net *n*-vectors output as a 3d plot

capabilities, we train a Ge3Net network using ancient genomes from the time period of 1700-3000 BP (Before Present). During

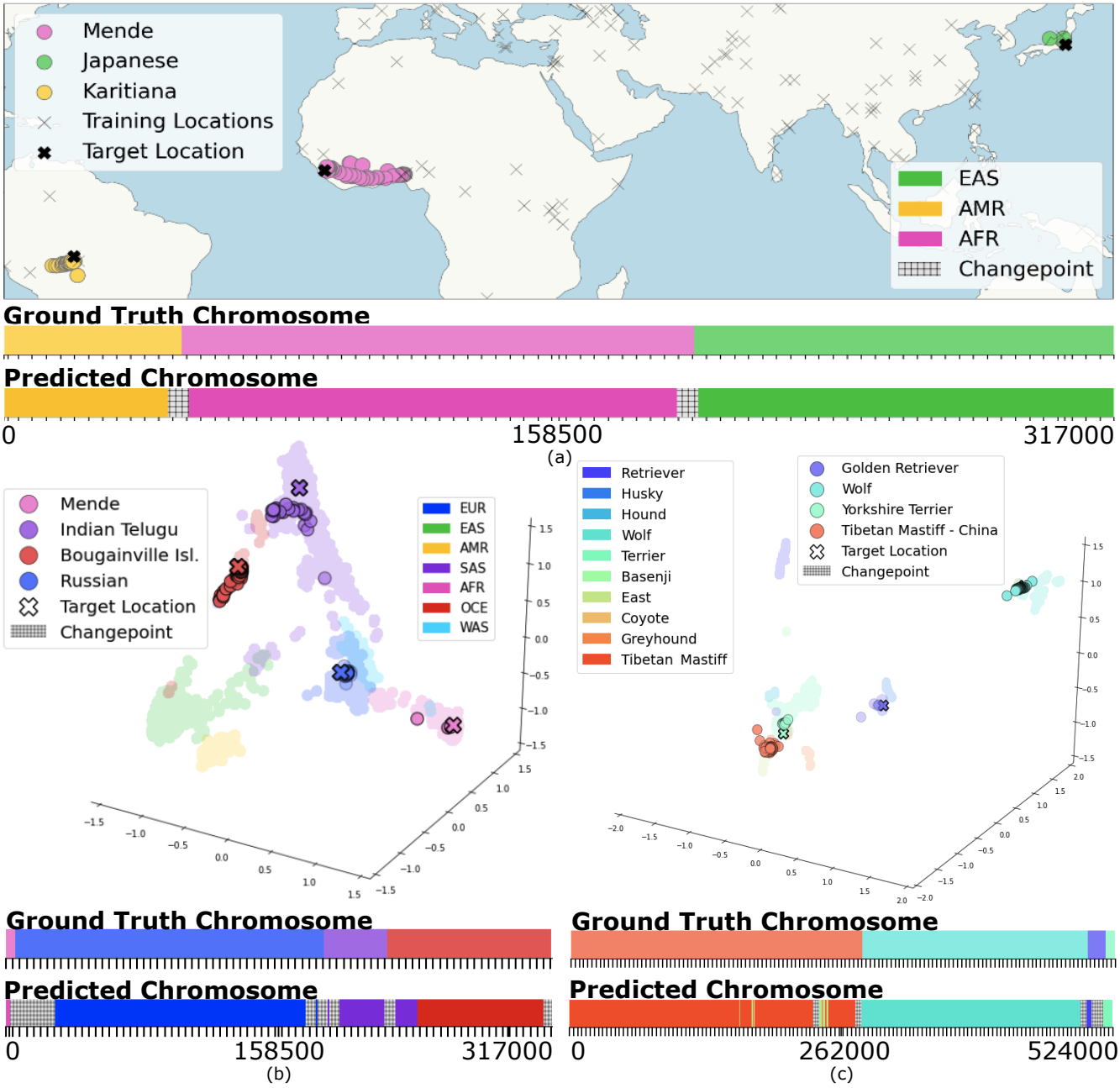

Figure 2. Continuous Ancestry Inference for Simulated Admixed Haplotypes a) Geographical Inference for a simulated admixed haplotype for humans **b)** Inference for a simulated admixed haplotype over PCA generated space for humans **c)** Inference for a simulated admixed haplotype over UMAP generated space for canid species

training, only sequences from chromosome 1 with coverage > 1 are used.

Figure 3 (a) - (b) shows inference of modern HGDP Kailash Haplotype on Ge3Net trained with ancient samples on the left and the corresponding sample predictions from Ge3Net trained on modern samples on the right. Figure 3 (c) shows the predictions of an ancient Turkish sample dated 3800 BP and (d) shows the shows the predictions of a modern Turkish sample, dated 375 BP , both inferred on the same Ge3Net trained with ancient samples belonging to the time period of 1700-3000 BP. Finally, Figure 3 shows the predictions for an ancient haplotype dated 5199 BP referred to as Iceman with Ge3Net model trained with ancient samples in the time period of 1700-3000 BP. We note that generally the inference from Ge3Net ancient shows traces of ancestry that are historically consistent. The chromosome plot shown with each geographical plot uses a latitude and longitude gradient scale. We plot the predicted coordinates according to this scale, disregarding continental or any other

boundaries.

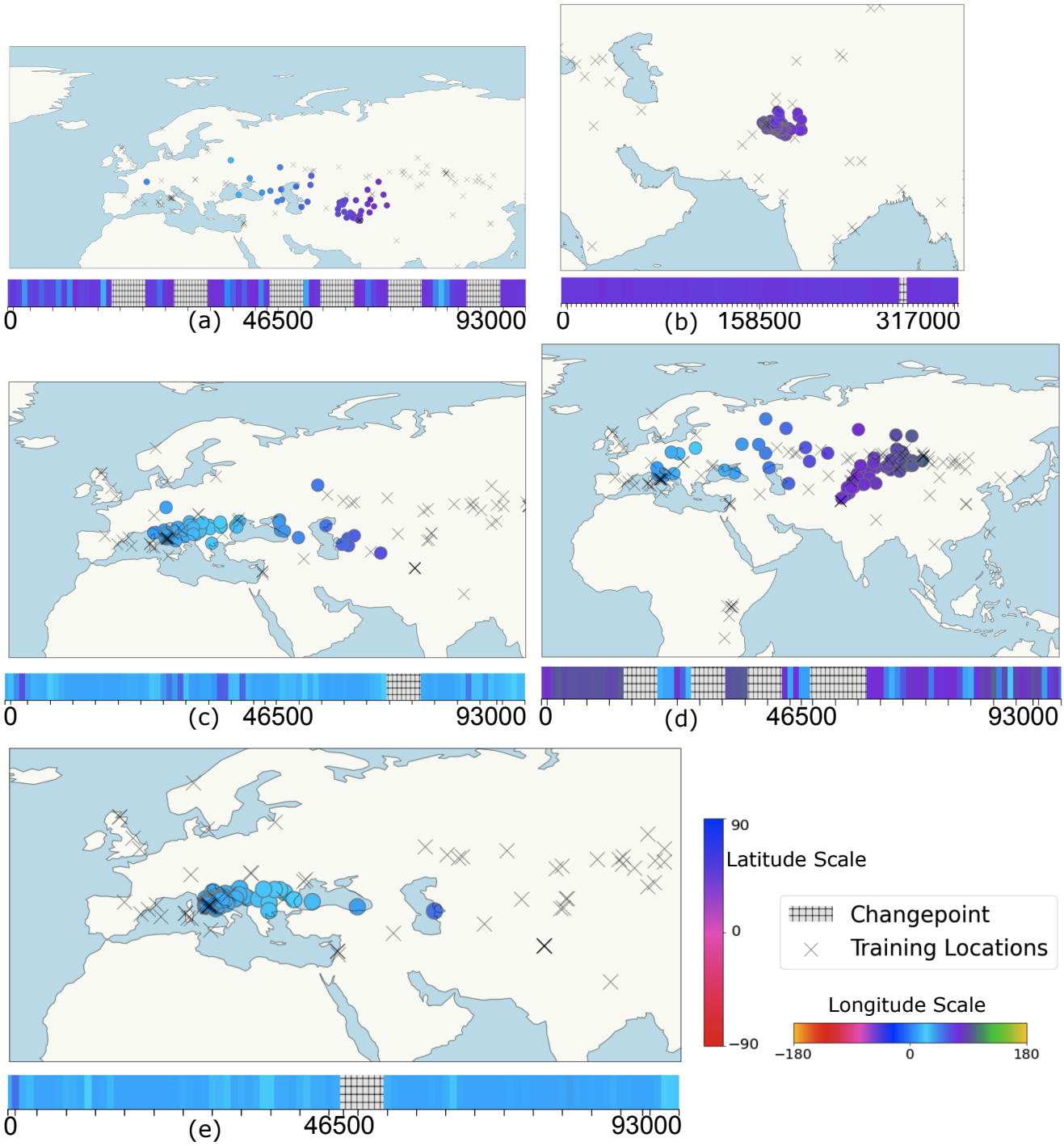

Figure 3. Continuous Coordinate Inference for Ancient Haplotypes (a)-(b) Kailash Haplotype with sample ID **HGDP00279** with model trained on ancient samples and with model trained on modern samples **(c)-(d)** Turkish ancestry ancient haplotype on left with sample ID **MA2200_final.SG**, dated **3800 BP** and a modern haplotype on right with sample ID **MA2195_final.SG**, dated **375 BP** of **(e)** Haplotype with sample ID **Iceman.SG**, dated **5199 BP** referred to as Iceman

2.5 Geometric Regression

We perform SVD decomposition of the snps for human genotype and project them to n dimensional space (3 principal components as they explain 8% of the variation). In order to ensure that the components on each axis are normalized, we perform whitening. This space is meant to provide global coordinates and for that reason, we combine the snps across all 22 chromosomes and apply a MAF of 0.1. A total of 5.8 million snps are used to form this space. These are fed as target coordinates for Ge3Net and we find that the structure recovered from predictions matches that of the original space. The coordinates are unique for each haplotype and represent the realistic scenario where ancestry is not binned but rather a continuous system of coordinates. This allows us to differentiate between the coordinate of individuals that have been assigned the same label and instead form the coordinate system by a method (PCA in this case) that captures the genetic variation. This approach allows for high resolution ancestry coordinate regression without the knowledge of labels. In this way, our method allows for any linear or non-linear method such as UMAP, t-SNE, auto-encoders to be easily adapted into the pipeline and used for local ancestry coordinate regression. Figure 8 show that we are able to recover the overall structure from Ge3Net predictions using either geography, or unsupervised method - PCA and UMAP generated spaces. For Canid species, we combine chromosomes 1 through 38, resulting to a total of 19.40 snps to be used for constructing the space. The impact of Linkage Disequilibrium (LD) on PCA has been well studied[40,41] and there is evidence that Canids have mucher higher levels of LD present in comparison to humans ($\sim$ 2 Mb in the canids compared with $\sim$ 0.28 Mb in Homo sapiens)[42,43] because of which the first few principal components do not capture the variation in many of the canid species. We use non-parameteric UMAP to find the global coordinates, which are used as target coordinates to train single ancestry dog breeds. The predictions from Ge3Net for admixed dog breeds reveal historical patterns as shown in Figure 9. The African Village Dog clearly shows segments close to Basenji. Similarly Vietnam Indigenous Dog shows segments that are most related with the East Asian and Tibetan Mastiff Canid single ancestry species.

2.6 Experimental results

Table 1 shows the mean balanced GCD for different architectures. Since the task of inferring coordinates for local ancestry is novel, we formulate a benchmark with LAI-Net (local ancestry inference for classification based on neural networks) by taking the granular populations (136-way classification) predicted by the method and computing GCD between the predicted granular population location and the labeled granular pop location. This method is not perfect since the GCD will be ~ 0 for populations the method correctly predicted. Detailed comparison between Ge3Net and LAI-Net is shown in 9. We created internal benchmarks gradually increasing in more complex non-linear transformations. We compare with 1) MLP layers that are window specific Fully Connected (FC) layers, 2) MLP window specific layers for Base module with an auxiliary network also consisting of MLP layers, 3) MLP for base and auxiliary network and Conv1D as a smoother similar to LAI-Net architecture, 4) MLP for base network without any auxiliary network and Attention modules as a smoother , 5) MLP for base and auxiliary network and BiLSTM as a smoother , 6) MLP for base and auxiliary network and Attention modules as a smoother, 7) MLP for base and auxiliary network and a single Attention head followed by 2 layers of BiLSTM as a smoother.

Table 1. Mean Balanced Great circle distance (Km) $\downarrow$

Network	Validation	Test
LAI-Net (Large)	-	1112.19
LAI-Net (Small)	-	1501.21
MLP	2142.32	2082.29
MLP (Base/Aux)	2164.52	2097.45
MLP + LSTM	2002.83	1976.26
MLP (Base/Aux) + Conv1D (Smoother)	1399.60	1355.36
MLP (Base/Aux) + Attention Head (Smoother)	1080.11	1060.74
MLP (Base/Aux) + LSTM (Smoother)	1020.13	999.63
MLP (Base/Aux) + Attention Head (Smoother)	939.14	913.30
MLP (Base/Aux) + Attention Head + LSTM (Smoother)	928.07	902.40
Ge3Net (MLP (Aux) + 2 Attention Heads + LSTM)	**866.68**	**828.57**

2.7 Noisy Labels And Hidden Structure

The ancestry labels assigned to each individual in the training set are based on geographic origin, but due to large-scale migrations that affected a few of these populations some of their genomic segments will have originated from geographically distant ancestries in the near past. Thus, some training individuals will contain genomic segments with geographic origins far

from the individual's labeled origin. For example, Figure 4 (a) shows an example of Ge3Net prediction for an individual with Hazara ancestry, showing chromosomal segments originating from Afghanistan, the country in which the Hazara live, as well as closer to Mongolia. This is consistent with the Hazara's history as having partially descended from the medieval Mongol conquerors of Afghanistan[44]. Figure 4 (b) shows Ge3Net prediction for a sample originating from Brahui and confirms that this sample is single ancestry. Figure 4 (e) shows inference for a haplotype labeled as an African-American. We observe that the predictions show a continuum of European (British) and African ancestries. These examples shows that Ge3Net can help to detect structure within populations.

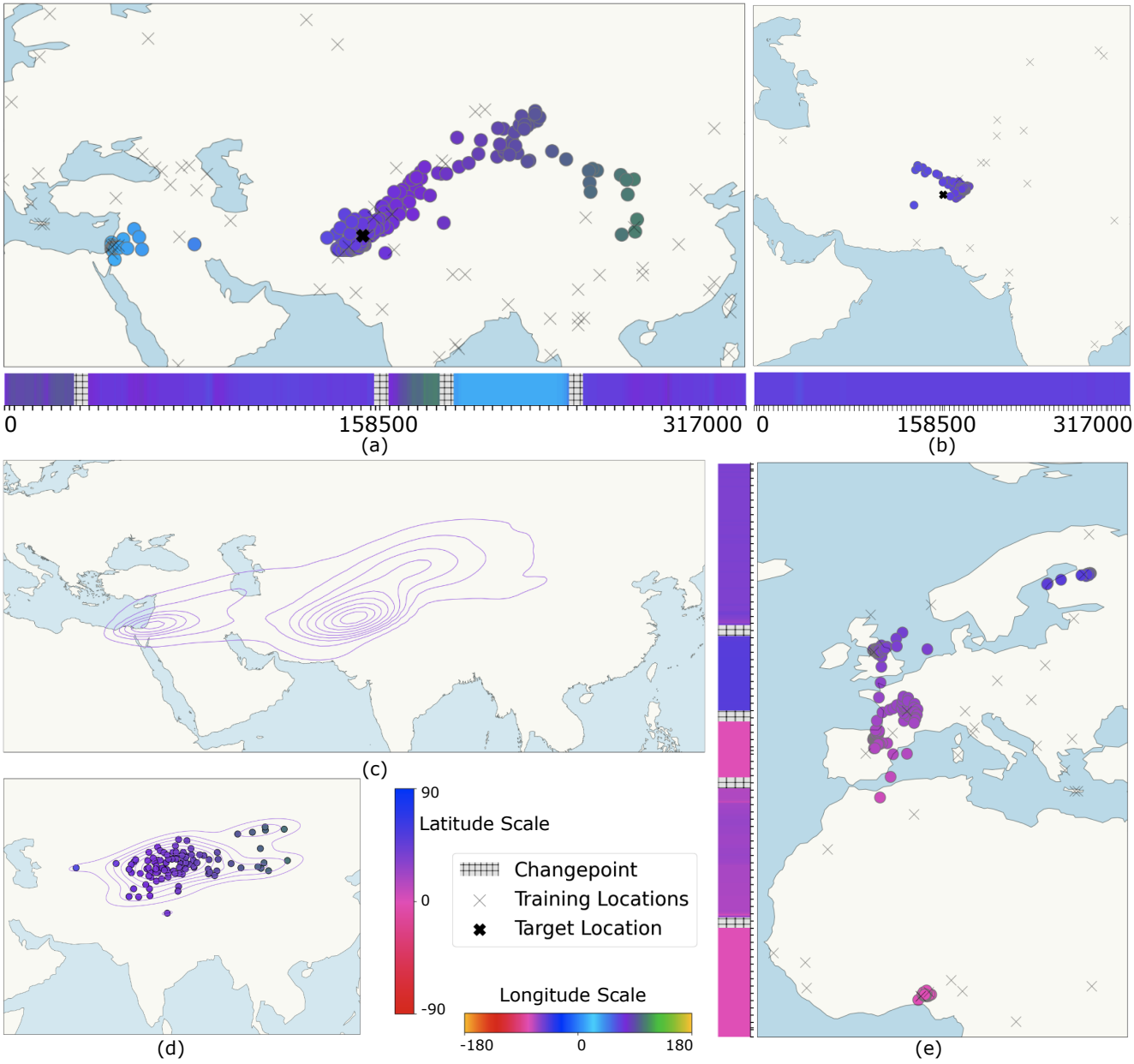

Figure 4. Real Admixed Haplotypes (a) Continuous Ancestry Inference for one of the Hazara Haplotype with Sample ID **HGDP00122 (b)** Continuous Ancestry Inference for one of the Brahui Haplotype with Sample ID **HGDP00025 (c)** Kernel Density Estimate for one of the Hazara Haplotype with Sample ID HGDP00122 **(d)** Kernel Density Estimate for a single window with 100 sampled predictions for pne of the Hazara Haplotype with Sample ID HGDP00122 **(e)** Continuous Ancestry Inference for one of the African-American Haplotype with Sample ID **NA20126**

2.8 Monte Carlo Dropout For Uncertainty Quantification

Being able to quantify the uncertainty and confidence of predictions is a key feature for a model designed for biomedical applications. By applying Monte Carlo Dropout (MC Dropout)[45] we can estimate the variance of the predictions for each window. Furthermore, by sampling multiple coordinates within (and among) windows, we can obtain a point cloud that, when combined with a kernel density estimator (KDE), provides a probabilistic representation of the individual's geographic ancestry. Figure 4(c) shows the estimated density of the geographic ancestry predictions from a genetic sequence sequence using a Gaussian KDE (with bandwidth=0.05, selected via cross-validation with the validation samples) for Hazara sample HGDP00122. The peaks of the estimated density are seen to be close to the ground-truth locations. Figure 4 (d) shows the 100 samples prediction for a specific window along with the Gaussian KDE point cloud for the particular window. By analyzing regions with high error or uncertain predictions, historical migrations can be unveiled.

2.9 Changepoint Detection

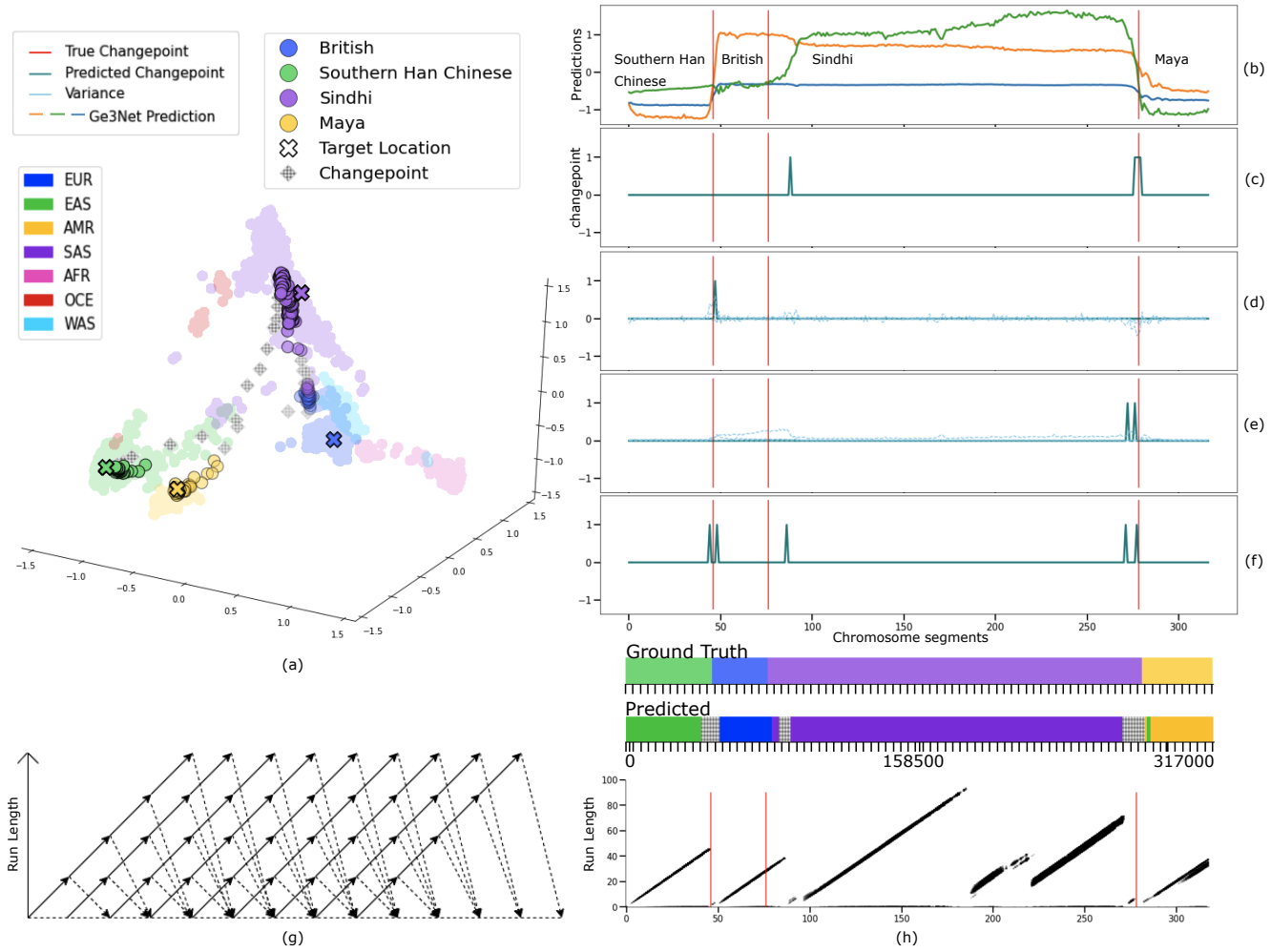

Figure 5. Changepoint Detection from Ge3Net Predictions (a) Continuous Ancestry Inference for a simulated admixed haplotype with detected changepoints **(b)** Ge3Net predictions with true changepoints **(c)** Neural network detected changepoints **(d)** Changepoints detected by change in the predictions **(e)** MC Dropout detected changepoints **(f)** BCD detected changepoints **(g)** Conceptual diagram of change in probabilities of BCD **(h)** The joint distribution obtained from BCD for the simulated admixed haplotype

With admixed haplotypes, it is critical to distinguish the breakpoint at which ancestry changes. For coordinate regression, this can be quite complex as the ancestry coordinates are continuous and can gradually transition from one ancestry to the next. The changepoints are also areas within the haplotype where gcd error is the highest and thus intuitively are source of uncertainty. Thus, changepoint detection is critical to our understanding of whether a predicted coordinate should be trusted as model error due to mixed signal at transition points or as interesting patterns of real migration affects of population. During training, the

changepoints and ± 1 window on either side of the changepoint are masked, meaning that loss is not propogated for these windows. This means that during inference, the changepoints give high errors and thus are easier to get detected as outliers.

For detecting changepoints, we explore the following methods -

- Neural network based changepoint detection: We find that using a branch from the 64 dim representation to predict a binary for changepoint with the additional Binary Cross Entropy as loss improves the network prediction. Since there is imbalance (lower number of positives- changepoints), we use the scaling hyperparameter of the binary cross entropy loss to upsample the changepoints.
- Uncertainty from MC Dropout based changepoint detection: In order to capture the variance in predictions with MC Dropout, we capture 100 samples with dropout enabled for non-normalizing layers and detect changepoints based on threshold (as a hyperparameter) of the change in variance of these samples predictions per segment.
- Gradient of n-vectors: Detecting changepoints based on threshold (as a hyperparameter) of change in n vector. While this approach is simple and works well, it fails for cases when the change in n vector is gradual versus a drastic.
- Bayesian Changepoint detection (BCD): This approach involves applying Bayesian Online Changepoint Detection[46] based on product partition models[47]. The method predicts the posterior distribution of run length r_t, where run length is defined as the as the continuous set of genomic segments since the last changepoint. At any given point across the genome, run length r_t keeps on monotonically increasing , except at changepoints, where it drops. The underlying probabilistic model predictive $p(x_{t+1}|r_t, x(l)) = \int p(x_{t+1}|\eta)p(\eta|x_{1:t}, r_t)\, d\eta$ where η refers to the parameters of the Gaussian distribution (we assume that the n-vectors for a particular cluster/ancestry form a gaussian distribution and is different than the gaussian of the next cluster/ancestry). x is the observable n vector values at each window. The method uses bayesian inference of Gaussian with conjugacy of exponential family, since we are considering Gaussian distribution and update the sufficient statistics, to obtain the posterior distribution $p(r_t|x_{1:t})$. Figure 5 (g) shows the conceptual idea[46] that at any given point the posterior probability either increases, resulting in increase in run length or that the probability mass goes down resulting in changepoint. The corresponding run length for a simulated admixed sample is shown in Figure 5 (h). Note that there is no changepoint detected by BCD (Figure 5 (f)) between window segments 100 and 250, even when the run length changes between that segment. This is because the run length drops down to a minimum of 6.0 between these window segments, indicating not a significant change in probability mass.

Figure 5 shows the comparison of different methods for a simulated admixed haplotype with four true changepoints. While none of the method perfectly predict all the four changepoints, BCD performs best and is able to detect the four changepoints within a few window segments of the true changepoint. The second changepoint is paticularly hard to detect between British and Sindhi ancestry. The similarity between South East Asian and European ancestries make it particularly challenging. The methods relying on variance of predictions or variance of uncertainty (Figure 5 (d) and (e)) fail as the change is gradual. Neural network and BCD are able to detect with the gap of some window segments. We also explore the correct evaluation metric to analyze changepoint detection and use the following regime - given a set of changepoints for a haplotype, we consider that it is a true positive if the detected changepoint is within the tolerance of true changepoint.Example: True changepoints = [3, 31, 250, 273], predicted changepoints = [30, 31, 249, 250, 272, 300], then True positives = 3, False positives = 1, False negatives = 1, True Negatives = 312 since there are a total of 317 chromosomal segments, out of which changepoints 31, 250 and 273 were detected within the allowed window tolerance of ± 1. The true changepoint 3 could not be detected and false changepoint 300 was detected. The performance for different method over different metrics is shown in Table2, where the hyperparameter for each method is chosen over a validation set w.r.t F1-Score and metrics reported for a held-out test dataset. We find that BCD is robust across different coordinate systems. In order to further analyze changepoint detection, we evaluate GCD error on test dataset in four different ways as shown in Table 3. Let cp denote true changepoints and $\hat{cp}$ denote the predicted changepoints and GCD between n-vectors as described in 4.2. No Masking refers to $\text{GCD}_{n_{x,y,z},\hat{n}_{x,y,z}}$ without removing either cp or $\hat{cp}$. True Changepoint Masking refers to $\mathbb{1}(n_{x,y,z} \neq cp)\text{GCD}_{n_{x,y,z},\hat{n}_{x,y,z}}$, Intersect Changepoint Masking refers to $\mathbb{1}(n_{x,y,z} \neq (cp \& \hat{cp}))\text{GCD}_{n_{x,y,z},\hat{n}_{x,y,z}}$ and Predicted Changepoint Masking refers to $\mathbb{1}(n_{x,y,z} \neq \hat{cp})\text{GCD}_{n_{x,y,z},\hat{n}_{x,y,z}}$. As expected, we observe lower gcd when conditioned with changepoints.

3 Discussion

A geographic regression-based local ancestry algorithm is demonstrated for the first time using a transformer based framework: Ge3Net. The continuous outputs of this method give high-resolution ancestry predictions along the genome and around the globe, eliminating the need for a multiplicity of ill-defined ethnic categories (and idealized ethnic references) used by traditional classification-based methods. Ge3Net benefits from diverse training data spread across the globe and allows for interpolation between ancestries in a continuous fashion, which is particularly valuable in regions that lie between traditional ancestry

Table 2. Evaluation of different Changepoint Detection Methods

Method	Coordinate System	Precision	Recall	Balanced Accuracy	Accuracy	F1-score
Neural Networks	PCA	65.82	60.34	79.90	98.80	62.96
Gradient change	PCA	73.84	49.27	74.49	98.85	59.10
MC Dropout	PCA	37.05	42.89	70.82	97.80	39.76
BCD	PCA	61.69	64.47	**81.89**	98.72	**63.05**
Neural Networks	Geography	16.70	63.75	80.52	97.0	26.46
Gradient change	Geography	48.92	61.75	80.60	99.13	54.59
MC Dropout	Geography	39.58	45.26	72.34	98.95	42.23
BCD	Geography	63.20	61.53	**80.61**	99.37	**62.35**

Table 3. Mean Balanced GCD (Km) ↓ Evaluation with various changepoint masking regimes

No Masking	True Changepoint Masking	Intersect (True w/ Pred) Changepoint Masking	Predicted Changepoint Masking	generation
588.81	588.81	588.81	586.39	0
735.84	718.43	722.87	728.38	2
871.57	838.55	847.97	858.95	4
1040.13	979.63	997.05	1018.13	8
828.57	798.38	806.89	816.67	all

classes (e.g. North Africa, between Sub-Saharan Africa and Europe, and Central Asia, between South Asia and East Asia and Europe). Ge3Net produces continuous ancestry coordinates for each genomic segment that can be used as covariates for genetic association studies and polygenic risk scores. In this way, Ge3Net could be used to capture ancestry-specific linkage effects, thus aiding in building more accurate genetic prediction models for admixed individuals.

A limitation of our current approach of inferring ancestries along a continuous spectrum as compared to continental classification is the difficulty to infer ancestries for very high generation of admixed individuals. For example, for simulated admixed individuals, we train the model for generations up to 32 and the test set consists of generations up to eight. Future work could explore effective methods to infer high-generation admixed ancestries. A related future direction is to explore the tension between inferring small segments with high confidence versus the interpolated erroneous segments at ancestry switches.

This work targets the disparities in genomic medicine that are emerging between European-descent populations and all other populations. By annotating ancestry along the genome accurately, and with simple to use coordinates, we hope to enable genetic researchers to incorporate ancestry-specific genetic effects into their future models with ease. This could help to extend the benefits of such research and models to more diverse cohorts. Ge3Net is particularly targeted at improving genetic modeling applied to admixed individuals. Such individuals inherit genomic segments from diverse populations that have very different genetic correlations (linkage). This ancestry-specific structure must be identified for each segment of the genome to apply appropriate ancestry-specific risk models.

4 Methods

4.1 Model Architecture

We use deep neural networks - specifically, we use a transformer encoder with raw allelic sequences for admixed population as input. A Bi-directional LSTM layer is used after the transformer encoder, followed by a fully connected layer to predict 3 points (n Vector) for the geographical task. Additionally, a small fully connected network with 3 layers interspersed with layernorm and relu activation is used for the prediction of changepoints, using a binary cross entropy loss. Further, a small auxiliary network consisting of 2 fully connected layers interspersed with relu and layernorm is used for the same regression task as the main task with GCD as the loss. In this way, the network performs multi-task operation with two different tasks - a regression task between the predicted and labeled n-vectors for each window (also referred to as a genomic segment with the window length of 1000 snps used as a hyperparameter in all experiments) and a binary task of predicting changepoints where ancestry switches are predicted.

4.2 Training Setting

Key to our method for the geographical coordinate system, is the concept of optimizing for n-vectors[48] as opposed to latitude and longitude. This ensures there is no discontinuity, for example, when longitude switches from Alaska to Russia. Training with n-vectors also helps to prevent distortion of distances at high latitudes, for example - near poles. n-vectors are unit vectors along three dimensions and conversion of latitude and longitude to n-vectors is defined as $n_x = \cos(\text{lat})\cos(\text{long})$, $n_y = \cos(\text{lat})\sin(\text{long})$, and $n_z = \sin(\text{lat})$. The latitude and longitude can thus be obtained from $\text{lat} = \arcsin(n_z)$ and $\text{long} = \arctan(\frac{n_y}{n_x})$. After the transformation from latitude/longitude of labels to n-vectors as defined above, Great Circle distance between labeled $n_{x,y,z}$ and predicted $\hat{n}_{x,y,z}$ n-vectors is defined as $\text{GCD}_{n_{x,y,z},\hat{n}_{x,y,z}} = \arccos(n_{x,y,z} \cdot \hat{n}_{x,y,z})$*Earth's Radius We train with the higher generation (gen) of simulated admixed as compared to the validation or test set so that the network is trained for a much harder task than it will be evaluated on. Specifically, we train for gen 0,2,4,8,16,24 and evaluate for gen 0,2,4 and 8.

4.3 Objective Function

For the geographical coordinate system, the loss used is the sum of the Great Circle Distance for the normalized n-vectors for both the Main and Auxiliary branch of Ge3Net. We use a masked loss where the transition windows, also referred to as changepoints are masked. This makes the network learn the pattern of changepoints. As a result, for inference, Ge3Net predicts a stark difference in the n-vectors for the change in inferred ancestry. For training Ge3Net with the geographical coordinate system, the objective function is given by

$$\mathcal{L}_{Geography} = \mathbb{1}\{n_{x,y,z} \neq \text{true changepoint}\} \cdot (GCD^{\text{Main}}_{n_{x,y,z},\hat{n}_{x,y,z}} + GCD^{\text{Aux}}_{n_{x,y,z},\hat{n}_{x,y,z}})$$

For PCA or UMAP based coordinate system, the objective function is given by

$$\mathcal{L}_{PCA/UMAP} = \mathbb{1}\{c_{x,y,z} \neq \text{true changepoint}\} \cdot (L1^{\text{Main}}_{c_{x,y,z},\hat{c}_{x,y,z}} + L1^{\text{Aux}}_{c_{x,y,z},\hat{c}_{x,y,z}})$$

where L1 loss between the true coordinate $c_{x,y,z}$ and predicted coordinate $\hat{c}_{x,y,z}$ is computed analogously for both the main and auxiliary network.

4.4 Dataset Preparation

We perform our experiments on three types of genotypes and discuss genotype specific processing in 4.4.1, 4.4.2, and 4.4.3. In general, we take the publicly available vcf files, and perform pre-processing including phasing and imputation when necessary.

4.4.1 Human Dataset

We use a dataset composed of the most comprehensive set of whole genome sequences collected for public use from real worldwide populations; namely, the 1000 genomes project[49], the Simons Genome Diversity Project[50], and the Human Genome Diversity Project[51] each pruned to include only single-ancestry origin individuals (see appendix A). Because the ground-truth ancestry switch points along the genome are not known for true admixed individuals, we used these single-ancestry individuals to simulate admixed descendants of varying generations, using a standard recombination model based on the human HapMap genetic map[25]. These descendants are simulated only insofar as the recombination (mixing) points of their parental ancestries are chosen at random; their parental genome sequences originate from our real genome sequence dataset. From 3156 real single-ancestry individuals, we selected 2020 to generate 4040 admixed descendants for training, 568 to generate 1800 admixed descendants for validation, and the remaining 568 to generate 1800 admixed descendants for testing. We have used Chromosome 22 for our analysis, but the results easily transfer to other chromosomes.

4.4.2 Canid Dataset

We have used Chromosome 22 for our analysis, but the results easily transfer to other chromosomes.

4.4.3 Ancient Dataset

We process the raw ancient genomic data by converting ANCESTRYMAP format to PACKEDPED format using open source software CONVERTF[52,53]. Using another open source software PLINK[54], we process the genomic files into individual chromosomes, phase and impute the missing snps using Beagle. We then divide the dataset into three time slots - samples that are very ancient before 3000 BP, 1700-3000BP, and modern samples from current-1700BP. We simulate admixed individuals for the second time slot 1700-3000BP and do inference on much more ancient (samples from the first time slot) or modern (samples from the third time slot). Since ancient haplotypes have generally low coverage, we use Chromosome 1 to extract as much genotype information.

5 Data & code availability

The code to reproduce results, and train models is available at https://github.com/RichRast/Ge3Net. The input datasets for our human genome modeling and analyses are available publicly through the Human Genome Diversity Project (HGDP), Simons Genome Diversity Project (SGDP), and 1000 Genomes Projects. Data for the canid genomes is available through NCBI and for ancient data through the Ancient Genome Diversity Project AGDP

6 Competing interests

The authors declare no competing interests.

7 Acknowledgments

Computing for this project was performed on the Sherlock cluster at Stanford University. We would like to thank the Stanford Research Computing Center for providing computational resources and support.

References

1. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104, DOI: 10.1126/science.1153717 (2008).
2. DeGiorgio, M., Jakobsson, M. & Rosenberg, N. A. Out of Africa: modern human origins special feature: explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc. Natl. Acad. Sci. United States Am.* **106**, 16057–16062, DOI: 10.1073/pnas.0903341106 (2009).
3. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591, DOI: 10.1038/s41588-019-0379-x (2019).
4. Lehmann, B., Mackintosh, M., McVean, G. & Holmes, C. Optimal strategies for learning multi-ancestry polygenic scores vary across traits. *bioRxiv* DOI: 10.1101/2021.01.15.426781 (2022). https://www.biorxiv.org/content/early/2022/04/07/2021.01.15.426781.full.pdf.
5. Mathieson, I. The omnigenic model and polygenic prediction of complex traits. *Am. J. Hum. Genet.* DOI: 10.1016/j.ajhg.2021.07.003 (2021).
6. Rajabli, F. *et al.* Ancestral origin of ApoE ε4 Alzheimer disease risk in Puerto Rican and African American populations. *PLoS Genet.* **14**, e1007791, DOI: 10.1371/journal.pgen.1007791 (2018).
7. Wang, S. *et al.* Genetic variants demonstrating flip-flop phenomenon and breast cancer risk prediction among women of african ancestry. *Breast cancer research treatment* **168**, 703–712 (2018).
8. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nat. News* **538**, 161–164, DOI: 10.1038/538161a (2016).
9. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
10. Hsu, P., Bryant, M. C., Hayes-Bautista, T. M., Partlow, K. R. & Hayes-Bautista, D. E. Racially ambiguous babies and racial narratives in the united states: A growing contradiction for health disparities research. *Acad. Medicine* **94**, 1099–1102 (2019).
11. Wills, M. Are clusters races? a discussion of the rhetorical appropriation of rosenberg et al.'s "genetic structure of human populations". *Philos. Theory, Pract. Biol.* **9** (2017).
12. Tang, H., Peng, J., Wang, P. & Risch, N. J. Estimation of individual admixture: analytical and study design considerations. *Genet. epidemiology* **28**, 289–301 (2005).
13. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–59 (2000).
14. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**, 1655–64, DOI: 10.1101/gr.094052.109 (2009).
15. Alexander, D. H. & Lange, K. Enhancements to the admixture algorithm for individual ancestry estimation. *BMC bioinformatics* **12**, 246, DOI: 10.1186/1471-2105-12-246 (2011).
16. Cabreros, I. & Storey, J. D. A likelihood-free estimator of population structure bridging admixture models and principal components analysis. *Genetics* **212**, 1009–1029, DOI: 10.1534/genetics.119.302159 (2019). https://www.genetics.org/content/212/4/1009.full.pdf.

17. Gopalan, P., Hao, W., Blei, D. & Storey, J. Scaling probabilistic models of genetic variation to millions of humans. *Nat. genetics* **48**, DOI: 10.1038/ng.3710 (2016).

18. Meisner, J. & Albrechtsen, A. Haplotype and population structure inference using neural networks in whole-genome sequencing data. *bioRxiv* DOI: 10.1101/2020.12.28.424587 (2021). https://www.biorxiv.org/content/early/2021/09/01/2020.12.28.424587.full.pdf.

19. Martin, A. R. *et al.* An unexpectedly complex architecture for skin pigmentation in africans. *Cell* **171**, 1340–1353 (2017).

20. Tang, H. *et al.* Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.* **79**, 1–12 (2006).

21. Sundquist, A., Fratkin, E., Do, C. B. & Batzoglou, S. Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome research* **18**, 676–682 (2008).

22. Price, A. L. *et al.* Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLoS Genet.* **5**, 1–18, DOI: 10.1371/journal.pgen.1000519 (2009).

23. Sankararaman, S., Sridhar, S., Kimmel, G. & Halperin, E. Estimating local ancestry in admixed populations. *The Am. J. Hum. Genet.* **82**, 290–303 (2008).

24. Durand, E. Y., Do, C. B., Mountain, J. L. & Macpherson, J. M. Ancestry Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution. *bioRxiv* (2014).

25. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The Am. J. Hum. Genet.* **93**, 278–288 (2013).

26. Mas Montserrat, D., Bustamante, C. & Ioannidis, A. LAI-Net: Local-Ancestry Inference With Neural Networks. *Proc. IEEE Int. Conf. on Acoust. Speech Signal Process.* (2020). Barcelona, Spain.

27. Borowiec, M. L. *et al.* Deep learning as a tool for ecology and evolution. *Methods Ecol. Evol.* (2021).

28. Mantes, A. D., Montserrat, D. M., Bustamante, C. D., Giró-i Nieto, X. & Ioannidis, A. G. Neural admixture: rapid population clustering with autoencoders. *bioRxiv* (2021).

29. Perera, M. *et al.* Generative moment matching networks for genotype simulation. *bioRxiv* (2022).

30. Montserrat, D. M., Bustamante, C. & Ioannidis, A. Class-conditional vae-gan for local-ancestry simulation. *arXiv preprint arXiv:1911.13220* (2019).

31. Romero, A. *et al.* Diet networks: thin parameters for fat genomics. *arXiv preprint arXiv:1611.09340* (2016).

32. Wang, Z. *et al.* Automatic inference of demographic parameters using generative adversarial networks. *Mol. ecology resources* **21**, 2689–2705 (2021).

33. Lewis, A. C. F. *et al.* Getting genetic ancestry right for science and society. *Science* **376**, 250–252, DOI: 10.1126/science.abm7530 (2022).

34. Yang, W.-Y., Novembre, J., Eskin, E. & Halperin, E. A model-based approach for analysis of spatial structure in genetic data. *Nat. genetics* **44**, 725–731 (2012).

35. Battey, C. J., Ralph, P. L. & Kern, A. D. Predicting geographic location from genetic variation with deep neural networks. *ELife* **9**, e54507 (2020).

36. Vaswani, A. *et al.* Attention is all you need. *CoRR* **abs/1706.03762** (2017). 1706.03762.

37. Song, G. & Chai, W. Collaborative learning for deep neural networks, DOI: 10.48550/ARXIV.1805.11761 (2018).

38. Karavani, E. *et al.* Screening human embryos for polygenic traits has limited utility. *Cell* **179**, 1424–1435 (2019).

39. Gravel, S. Population genetics models of local ancestry. *Genetics* **191**, 607–619 (2012).

40. Zou, F., Lee, S., Knowles, M. R. & Wright, F. A. Quantification of population structure using correlated snps by shrinkage principal components. *Hum. Hered.* DOI: 10.1159/000288706 (2010).

41. Sim, S.-C. *et al.* High-density snp genotyping of tomato (solanum lycopersicum l.) reveals patterns of genetic variation due to breeding. *PLOS ONE* DOI: 10.1371/journal.pone.0045520 (2012).

42. Wayne, R. K. & Ostrander, E. A. Lessons learned from the dog genome. *Trends Genet.* DOI: 10.1016/j.tig.2007.08.013 (2007).

43. Sutter, N. B. *et al.* Extensive and breed-specific linkage disequilibrium in canis familiaris. *Genome Res.* DOI: 10.1101/gr.3147604 (2004).

44. Zerjal, T. *et al.* The genetic legacy of the mongols. *The Am. J. Hum. Genet.* **72**, 717–721, DOI: https://doi.org/10.1086/367774 (2003).

45. Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *Proc. Int. Conf. on Mach. Learn.* 1050–1059 (2016).

46. Adams, R. P. & MacKay, D. J. C. Bayesian online changepoint detection, DOI: 10.48550/ARXIV.0710.3742 (2007).

47. Barry, D. & Hartigan, J. A. Product partition models for change point problems. *The Annals Stat.* **20**, 260–279 (1992).

48. Gade, K. A non-singular horizontal position representation. *J. Of Navig.* **63**, 395–417 (2010).

49. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68 (2015).

50. Mallick, S. *et al.* The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

51. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367** (2020).

52. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* DOI: 10.1038/ng1847 (2006).

53. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLOS Genet.* DOI: 10.1371/journal.pgen.0020190 (2006).

54. Purcell, S. *et al.* Plink: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* DOI: 10.1086/519795 (2007).

Author contributions statement

All authors conceived the experiment(s) and brainstormed conceptual ideas. R.R. conducted the experiment(s) and analyzed the results. RR, DMM, and AGI reviewed the manuscript.

A Dataset Details

Tables 4 to 6 detail the genomic sequences used with the combined 1000 Genomes, HGDP, and SGDP as described in Section 4.4.1. The size of the reference haplotypes by subgroup population as annotated by these dataset sources is provided for full transparency.

Table 4. Human Train Dataset

Groupings	Size	Groupings	Size	Groupings	Size	Groupings	Size
Train							
Japanese	105	Burusho	19	San	5	BantuHerero	2
Yoruba	104	Basque	18	Mexican-American	3	Australian	2
Tuscan	92	Biaka	18	BantuSouthAfrica	3	Armenian	2
Gambian Mandinka	90	Kalash	18	Colombian	3	Ami	2
Spanish	86	Mandenka	18	Yemenite Jew	2	Hungarian	2
Southern Han Chinese	84	Hazara	15	Piapoco	2	Georgian Mingrelian	2
Indian Telugu	82	Adygei	13	Quechua	2	Igorot	2
Han Chinese	82	Orcadian	12	Relli	2	Khomani San	2
Gujarati	82	Karitiana	10	Saami	2	Mala	2
Sri Lankan	82	Bergamo Italian	10	North Ossetian	2	Madiga	2
Luhya	80	Pima	10	Saharawi	2	Luo	2
Kinh Vietnamese	79	Mbuti	10	Ulchi	2	Lezgin	2
Esan	79	Bougainville	9	Tajik	2	Kyrgyz	2
Finnish	79	BantuKenya	9	Thai	2	Kusunda	2
Punjabi	77	Uygur	8	Mixe	2	Korean	2
Dai Chinese	74	Yi	8	Tubalar	2	Iranian	2
British	73	Tu	8	Turkish Cappadocia	2	Zapotec	2
Bengali	69	NorthernHan	8	Yadava	2	Kapu	2
Mende	68	She	8	Abkhasian	2	Irula	2
Bedouin	37	Miao	8	Masai	2	Jordanian	2
Palestinian	37	Daur	8	Icelandic	2	Iraqi Jew	2
Druze	34	Oroqen	7	Greek	2	Albanian	1
Han	26	PapuanHighlands	7	Mansi	2	Altaian	1
Peruvian	25	Tujia	7	Even	2	Samaritan	1
Sardinian	22	Hezhen	7	Estonian	2	Atayal	1
Mozabite	22	Cambodian	7	Eskimo Sireniki	2	Polish	1
French	22	Xibo	7	Eskimo Naukan	2	Norwegian	1
Brahui	20	Dai	7	Dusun	2	Czech	1
Makrani	20	Mongolian	7	Dinka	2	Itelman	1
Russian	20	Naxi	6	Crete	2	Somali	1
Yakut	20	Maya	6	Burmese	2	Chane	1
Pathan	19	Lahu	6	Bulgarian	2	Eskimo Chaplin	1
Balochi	19	Surui	6	Brahmin	2	Khonda Dora	1
Sindhi	19	PapuanSepik	6	BantuTswana	2	Chechen	1

B Uncertainty Estimation

Figure 6 shows the multi-modal population structure of an Uygur haplotype. The inferred ancestry interpolates between East Asia and Central Asia represented by the interpolation of color (from green to purple) as depicted by the color scale according to geographical longitude Figure 6 subplot (a). We also show the kernel density plot for all the windows for chromosome 22 in Figure 6 subplot (b) representing the bimodal structure with one mode close to East Asia and the other close to Central Asia. Figure 6 subplot (c) shows the kernel density with 100 predictions from Ge3Net for a single window of chromosome 22.

Table 5. Human Validation Dataset

Groupings	Size	Groupings	Size	Groupings	Size	Groupings	Size
Validation							
Yoruba	13	Bedouin	5	Mbuti	2	Karitiana	1
Japanese	13	Druze	4	Orcadian	2	Hezhen	1
Tuscan	12	Han	4	Biaka	2	Lahu	1
Gambian Mandinka	12	Peruvian	3	Bergamo Italian	1	Maya	1
Han Chinese	11	Yakut	3	San	1	Even	1
Gujarati	11	Sardinian	3	She	1	Mexican-American	1
Spanish	11	Sindhi	3	Bougainville	1	Miao	1
Southern Han Chinese	11	Balochi	3	Tu	1	Mongolian	1
Kinh Vietnamese	10	Mozabite	3	Surui	1	Colombian	1
Luhya	10	Russian	3	Cambodian	1	Dinka	1
Finnish	10	Pathan	3	Tujia	1	Naxi	1
Esan	10	Makrani	3	BantuSouthAfrica	1	NorthernHan	1
Dai Chinese	10	Burusho	3	Uygur	1	Oroqen	1
Indian Telugu	10	French	3	Xibo	1	Daur	1
Sri Lankan	10	Brahui	3	BantuKenya	1	PapuanHighlands	1
Punjabi	10	Basque	3	Yi	1	Dai	1
Bengali	9	Adygei	2	Quechua	1	PapuanSepik	1
British	9	Kalash	2	Jordanian	1		
Mende	9	Hazara	2	Pima	1		
Palestinian	5	Mandenka	2	Mixe	1		

Table 6. Human Test Dataset

Groupings	Size	Groupings	Size	Groupings	Size	Groupings	Size
Test							
Yoruba	13	Mende	8	Brahui	2	Dai	1
Japanese	13	Bengali	8	Hazara	2	Cambodian	1
Tuscan	11	Druze	4	Kalash	2	Mbuti	1
Gambian Mandinka	11	Bedouin	4	Biaka	2	Pima	1
Esan	10	Palestinian	4	Basque	2	Daur	1
Kinh Vietnamese	10	Sardinian	3	Burusho	2	PapuanSepik	1
Indian Telugu	10	French	3	Lahu	1	PapuanHighlands	1
Han Chinese	10	Peruvian	3	Surui	1	Hezhen	1
Gujarati	10	Han	3	Bougainville	1	Oroqen	1
Finnish	10	Russian	2	Bergamo Italian	1	Orcadian	1
Luhya	10	Pathan	2	Tujia	1	NorthernHan	1
Sri Lankan	10	Sindhi	2	Tu	1	Naxi	1
Southern Han Chinese	10	Balochi	2	Uygur	1	Mongolian	1
Spanish	10	Mozabite	2	Xibo	1	Miao	1
Punjabi	9	Yakut	2	BantuKenya	1	Karitiana	1
Dai Chinese	9	Mandenka	2	Yi	1	Adygei	1
British	9	Makrani	2	She	1		

C Extended Ge3Net Results

We evaluate mean GCD and bin them by their respective labeled granular population as shown in Figure 7. We observe that the granular populations where GCD is higher are ancestries, for which there is historical evidence for migration, such as Uygur, Hazara, Xibo, Daur, and Yakut. Table 9 shows comparison between Ge3Net and LAI-Net for generations 0,2,4 and 8. Since LAI-Net was designed for the classification task, we report three metrics - GCD, continental classification, and granular pop

Table 7. Canid Dataset

Groupings	Size	Groupings	Size	Groupings	Size
Train					
Yorkshire Terrier	60	Scottish Terrier	3	Alaskan Malamute	2
Grey Wolf	22	Chow Chow	3	Iberian Wolf	1
Labrador Retriever	19	Siberian Husky	3	Scottish Deerhound	1
Golden Retriever	16	Basenji	3	Greenland Dog	1
West Highland White Terrier	9	Saluki	2	Chongqing Dog	1
Wolf	8	Shiba Inu	2	Tibetan Mastiff	1
Tibetan Mastiff - China	8	Afghan Hound	2	Cairn Terrier	1
Greyhound	7	Alaskan Husky	2	Whippet	1
Grey Wolf	6	Italian Greyhound	2	Xiasi Dog	1
Norwich Terrier	3	Coyote	2	Jindo	1
New Guinea Singing Dog	3	Chinese Shar-Pei	2		
Validation					
Yorkshire Terrier	8	Grey Wolf	1	Siberian Husky	1
Grey Wolf	3	Greyhound	1	Tibetan Mastiff - China	1
Labrador Retriever	3	Alaskan Malamute	1	West Highland White Terrier	1
Golden Retriever	2	New Guinea Singing Dog	1	Wolf	1
Chow Chow	1	Norwich Terrier	1	Afghan Hound	1
Coyote	1	Saluki	1		
Basenji	1	Scottish Terrier	1		
Test					
Yorkshire Terrier	7	Labrador Retriever	2	Tibetan Mastiff - China	1
Golden Retriever	2	Grey Wolf	1	West Highland White Terrier	1
Grey Wolf	2	Greyhound	1	Wolf	1

Table 8. Ancient Dataset (Time Range 1700-3000BP)

Groupings	Size	Groupings	Size	Groupings	Size	Groupings	Size
Train							
Pakistan	69	Cuba	8	South Africa	3	Norway	1
Mongolia	36	Latvia	7	USA	2	Malawi	1
Italy	35	Lebanon	7	Moldova	2	Laos	1
Kazakhstan	12	China	7	Nepal	2	Japan	1
Great Britain	10	Taiwan	4	Turkmenistan	1	Israel	1
Russia	10	Tanzania	3	Switzerland	1	Iran	1
Kenya	9	Hungary	3	Ukraine	1	Denmark	1
Spain	8	Kyrgyzstan	3	Bulgaria	1	Croatia	1
Venezuela	8	Vanuatu	3	Peru	1	Malaysia	1
Validation							
Pakistan	4	Italy	2	Mongolia	2	Kazakhstan	1

classification accuracy. We note that across all generations and metrics, Ge3Net performs better.

Figure 8 shows the inferred ancestry compared to the training locations for geographical predictions, PCA generated space for humans genotype and Umap generated space for canids. Overall, Ge3Net is able to learn and construct either the geographical space or the PCA/UMAP space based on the training dataset. This demonstrated the effectiveness and use of Ge3Net architecture for training on a user-defined space for different applications and genotypes.

Figure 9 shows inference for two real admixed canid species. We note that the model is correctly able to infer the Admixed

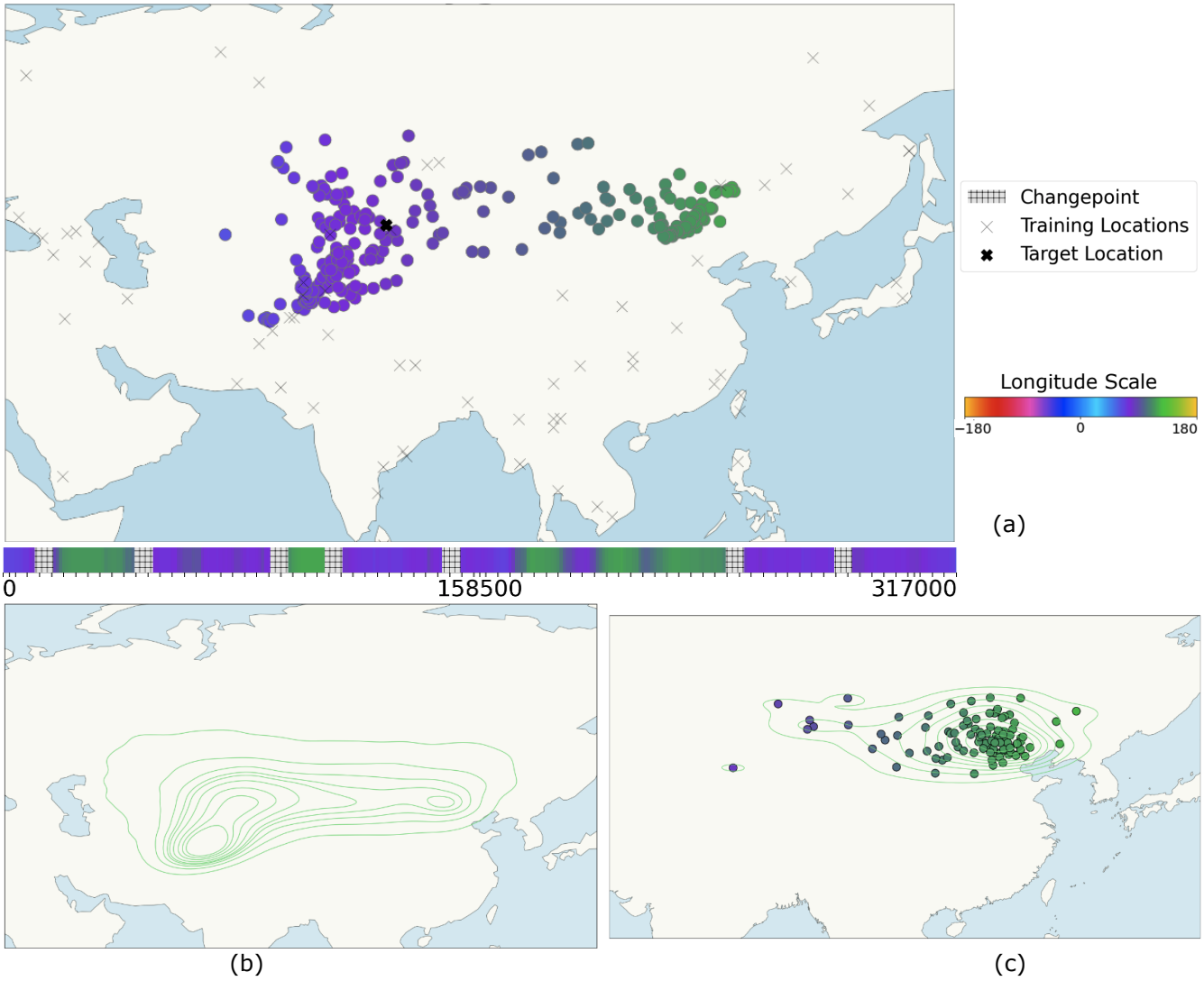

Figure 6. Mean GCD by granular population (a) Ge3Net predictions for an Uygur haplotype with Sample ID **HGDP01300** from chromosome 22 **(b)** KDE plot of Ge3Net predictions after MC Dropout for an Uygur haplotype from chromosome 22 **(c)** KDE plot of Ge3Net predictions after MC Dropout of a single window along chromosome 22 with 100 predictions

African canid species as interpolated with Basenji and Terrier and Vietnam canid species as most similar to the East/Tibetan Mastiff species that are present in the training dataset.

D Ge3Net HyperParameters

We use a batch size of 256 and learning rate of 1e-3 for all the layers except BiLSTM layers that use a learning rate of 1e-2. For optimization, AdamW optimizer is used and the model is trained for 300 epochs. Early stopping is used with the validation dataset if the loss keeps increasing or remains constant (change is detected at 1e-3 units for loss) for 6 consecutive steps. For the training objective, ± 1 window segments around the true changepoint are masked, so as to not to backpropagate and learn weights for those windows during training. For validation and testing, this masking is not done. Dropout of 0.3 for all the layers is used , except for Auxiliary network layers, where a dropout of 0.2 is used. Leaky Relu activation with a slope of 0.1 is used. Xavier initialization is used to initialize the weights of the network. A window size of 1000 is used for all the experiments in this paper. When using MC Dropout, 100 samples are generated per window for uncertainty estimates. For the changepoint neural network module during training, a scaling of 10.0 for binary cross entropy loss is used. For the changepoint methods described in 2.9, hyperparameters for each method are tuned on the validation dataset. For Neural network changepoint detection, the threshold at which logistic regression probability is used to detect a changepoint is treated as a hyperparameter. For changepoint

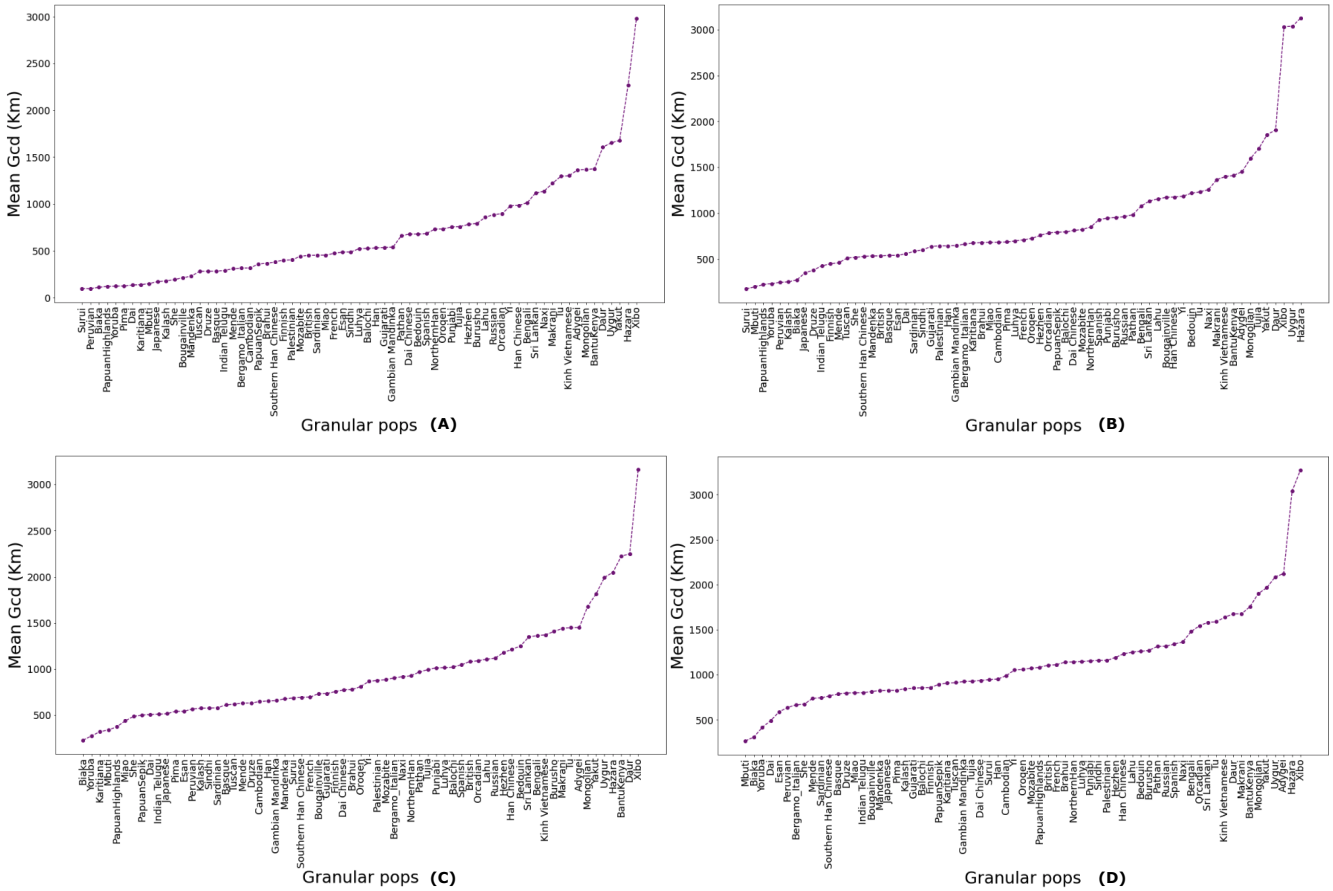

Figure 7. Mean GCD by granular population (A)-(D) Mean GCD by granular population for generation 0,2,4,8 respectively

Table 9. Classification Accuracy per generation

Network	Super Pop Classification Accuracy ↑	Granular Pop Classification Accuracy↑	GCD(Km) ↓	generation
LAI-Net (Small)	93.96	37.25	1281.96	0
LAI-Net (Large)	96.34	52.25	868.09	0
Ge3Net	97.36	46.76	**588.81**	0
LAI-Net (Small)	91.82	33.12	1465.61	2
LAI-Net (Large)	94.28	44.67	1103.66	2
Ge3Net	96.18	43.89	**735.84**	2
LAI-Net (Small)	90.61	30.92	1551.46	4
LAI-Net (Large)	93.43	43.23	1156.90	4
Ge3Net	94.62	37.61	**871.57**	4
LAI-Net (Small)	88.88	31.57	1634.54	8
LAI-Net (Large)	92.01	43.13	1240.79	8
Ge3Net	92.0	33.61	**1040.13**	8
LAI-Net (Small)	91.08	32.86	1501.21	all
LAI-Net (Large)	93.81	45.25	1112.19	all
Ge3Net	94.83	39.91	**828.57**	all

Table 10. Ge3Net Evaluation Metrics ↓

Metric	**Geography**	**PCA generated space**	**UMAP generated space**
Genotype	**Humans**	**Humans**	**Canids**
Mean Balanced GCD (Km)	828.57	-	-
Mean Balanced GCD by Super Pop (Km)	744.74	-	-
Mean Balanced GCD by Granular Pop (Km)	930.07	-	-
Median Balanced GCD (Km)	462.93	-	-
Median Balanced GCD by Super Pop (Km)	782.01	-	-
Median Balanced GCD by Granular Pop (Km)	793.13	-	-
L1 Loss	0.19	0.35	0.26
Smooth L1 Loss	0.02	0.10	0.06
MSE Loss	0.04	0.21	0.12

detection by gradient change, the threshold at which the variance of Ge3Net predictions exceeds a given value is chosen as a hyperparameter. For MC Dropout, the threshold at which the variance of 100 samples drawn exceeds a given value is chosen as a hyperparameter and for BCD changepoint detection, the value at which run length drops is chosen as a threshold. This value is generally selected as 3.0 based on the validation dataset. In addition, ± 3 windows within each predicted changepoint are also removed during visualization.

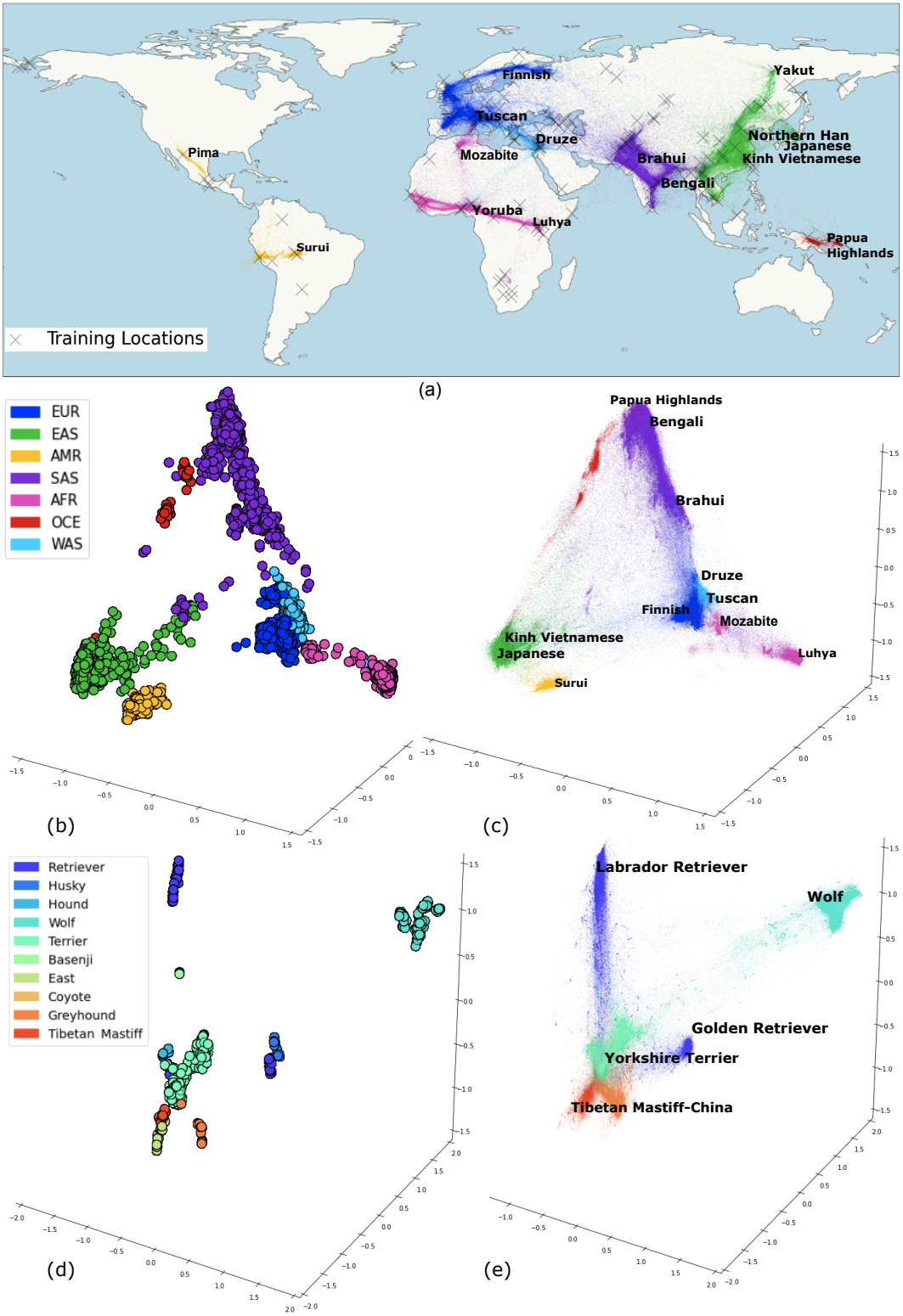

Figure 8. **Ge3Net Predictions versus Training Dataset a)** Geographical Predictions and Training Locations **b)** PCA generated space for humans representing training locations **c)** Ge3Net predictions over PCA space for humans **d)** UMAP generated space for canid species representing training locations **e)** Ge3Net predictions over UMAP space for canids

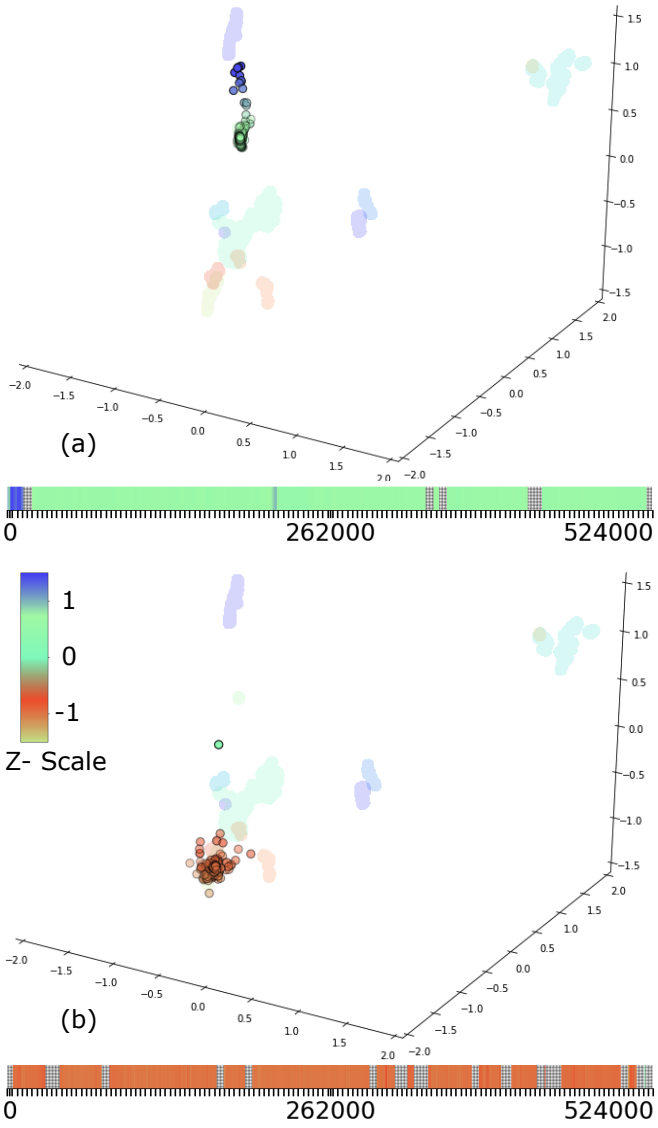

Figure 9. Ge3Net Predictions for Real admixed Canid Species

a) Continuous Coordinate Inference for real admixed haplotype over UMAP generated space for a Village Dog from Sub-Saharan Africa with sample name **VillDog_Africa01 b)** Continuous Coordinate Inference for real admixed haplotype over UMAP generated space for a Vietnamese Indigenous Dog with sample name **IndigenousDogVietnam04**