SNPScape: An Interactive Visualization System for Exploring Overlap and Uncertainty in Population Genetic Dimensionality Reduction

Jasmine C. Liu*, Alex Diaz-Papkovich, Sohini Ramachandran, and David H. Laidlaw

Abstract— Human genetic data is high-dimensional and complex, and population geneticists regularly use visualizations to both understand the data and present information in publications. However, common dimensionality reduction (DR) methods, such as principal component analysis (PCA) and uniform manifold approximation and projection (UMAP), often overemphasize discrete population separation and convey a sense of precise individual placement, making such visualizations susceptible to misinterpretation. We introduce *SNPScape*, a visualization framework for interpreting population-genetic embeddings using data from the 1000 Genomes Project that explicitly links feature-level diversity, embedding structure, and representational stability. *SNPScape* improves users' ability to derive insights into feature-driven clustering, embedding instability, and population overlap through integrating three coordinated views: (1) Euler diagrams summarizing overlap of common variants across populations, (2) PCA and UMAP embeddings generated from population-specific subsets of single nucleotide polymorphisms (SNPs), and (3) stability-aware overlays and contours that visualize positional variability of both superpopulations and individuals across repeated SNP subsampling. We evaluate the approach through a domain-expert study comparing the system against traditional static DR visualizations. Results show that *SNPScape* enables users to generate substantially more insights, perform more detailed multi-step reasoning, and interpret feature-driven structure and embedding uncertainty. These findings demonstrate the importance of uncertainty and feature-level contributions for understanding genetic visualizations, supporting more responsible interpretations and offering a new visualization framework for high-dimensional biological data.

Index Terms—Population genetics, uncertainty visualization, dimensionality reduction, PCA, UMAP, visual analytics, embedding stability.

1 Introduction

Dimensionality reduction (DR) visualizations are widely used in population genetics to visualize patterns of population structure and genetic ancestry [10, 11, 24, 25, 30]. In particular, PCA remains a standard analytical baseline in population genetics, while UMAP has increasingly been adopted for nonlinear structure exploration [23]. Despite their utility, both methods reduce high-dimensional genetic spaces into lower dimensions in ways that can be misleading and susceptible to misinterpretation [6, 12].

Two recurring interpretability risks motivate this work. First, static DR embeddings can overstate discrete separation between populations, encouraging interpretations of clusters as naturally distinct groups even when underlying genetic variation is continuous and overlapping. Second, embeddings are typically treated as stable outputs, despite their sensitivity to both the choice of genetic variants (e.g., SNP subsets) and stochastic variation in the embedding process. These factors can alter apparent structure, but are rarely made visible in standard population genetic visualization workflows. Existing tools expose either feature-level summaries or embedding-level views in isolation, leaving users without a direct, unified framework for reasoning about how feature selection, embedding geometry, and uncertainty interact.

We address this gap with *SNPScape*, a coordinated multi-view dashboard for uncertainty-aware exploration of population-genetic DR embeddings. *SNPScape* links: (i) overlap of common variants across populations, (ii) DR projections derived from population-unique SNP subsets, and (iii) stability overlays that quantify and display individual-level positional variance and contour overlays that show superpopulation-level distribution under repeated SNP subsampling and alignment. This design enables analysts to directly trace how variant selection influences observed structure and to assess the robustness of both group-level patterns and individual placements.

*Paper approved as a Master's report for this author.

• *Jasmine C. Liu, Alex Diaz-Papkovich, Sohini Ramachandran and David H. Laidlaw, are with Brown University. E-mail: jasmine_c_liu | alex_diaz-papkovich | sohini_ramachandran | david_laidlaw@brown.edu*

This paper makes the following contributions:

- **A unified visualization framework for feature-aware and uncertainty-aware population genetics analysis.** Unlike prior approaches that treat variant summaries, embeddings, or stability metrics in isolation, *SNPScape* integrates feature-level overlap, embedding structure, and uncertainty representations into a single interactive workflow, enabling analysts to directly trace how SNP selection influences projected structure.
- **A stability-driven visual encoding of embedding uncertainty.** We introduce a variance-based stability metric that quantifies representational ambiguity under SNP subsampling and integrate it with point-level markers and contour-based summaries, allowing users to identify unstable individuals and regions directly within embedding space.
- **Empirical evidence of improved analytical capability.** Through domain expert evaluations, we show that *SNPScape* enables increased insight generation and more accurate reasoning about feature-driven structure and embedding uncertainty compared to traditional PCA and UMAP visualizations.

2 Related Work

2.1 Dimensionality Reduction Visualization

Static PCA and UMAP [21, 23] plots remain the dominant baseline for population-structure visualization and are typically computed once from the full genotype matrix. This whole-dataset workflow is useful for high-level analysis, such as identifying broad gradients and major separations, but it offers limited visibility into which SNP groups contribute to the observed structure. Static PCA and UMAP views also omit explicit uncertainty measurements, such as the extent to which point locations and cluster boundaries change under SNP resampling or stochastic variation. As a result, analysts can overinterpret discrete boundaries and precise individual placement.

Our method preserves the interpretability benefits of PCA and UMAP while addressing these baseline constraints through coordinated interaction. Instead of treating embeddings as fixed endpoints,

SNPScape links projections to SNP overlap between populations and integrates uncertainty overlays, enabling users to evaluate both why a structure appears and how robust structures are under resampling. This integrated workflow supports richer task-specific insights, clearer detection of ambiguous individuals, and more responsible interpretation than static PCA and UMAP plots alone.

2.2 Population-Genetic Visualization Tools

Interactive visualization tools have long supported exploration of population genetic data, but typically focus on either variant-level summaries or low-dimensional embeddings in isolation. Kitchens and Coop [19] provide a high-level understanding of shared genetic diversity among populations through interactive Euler diagrams. The Geography of Genetic Variants (GGV) browser [22] provides an interactive map of allele-frequency distributions across populations, offering an intuitive, feature-centric view of shared and population-specific variation. Genome browsers and portals enable detailed exploration of variant annotations and population-level diversity statistics [5, 7]. However, none of these resources directly integrate these views with DR representations of structure.

SNPScape links feature-level overlap, embedding-level structure, and individual-level variability. In doing so, it connects diversity at the variant level with variability in low-dimensional representations, providing a multi-view framework for interpreting population structure as continuous, overlapping, and context-dependent rather than discretely partitioned.

2.3 Embedding Stability and Uncertainty

Uncertainty visualization has become increasingly important in the interpretation of DR [17, 26]. Prior work approaches this problem from two directions.

One line of work focuses on uncertainty arising from variability in the input data. In population genetics, stability is often assessed through feature resampling, producing both global and per-individual measures of robustness [14]. While these approaches provide important insights into feature-driven variability, they are primarily analytical and do not directly support visual interpretation of how instability manifests in embedding space.

A second line of work focuses on uncertainty introduced by the DR algorithms themselves. For non-linear methods such as UMAP and t-SNE, stochastic optimization and initialization can lead to variability in embeddings even when the input data is fixed. Visualization approaches in this space model algorithmic uncertainty by generating multiple embeddings or augmenting point representations to reflect variability across runs [18]. These methods provide fine-grained insight into embedding instability, but typically do not account for variability arising from changes in the underlying feature set.

Our work bridges these directions by jointly modeling feature-driven variability and embedding-level uncertainty within a unified visualization system. We introduce a variance-based stability metric derived from repeated SNP subsampling that captures representational ambiguity in DR embeddings. This metric is presented with stability-aware point representations, enabling users to directly interpret where and how variability arises. We also explore the robustness of embeddings through visualizing the distribution of points across the same random SNP subsamples, represented through contour-based summaries. By linking variant overlap, embedding structure, and stability within a single interactive workflow, our approach moves beyond isolated stability measures toward a more comprehensive and interpretable framework for uncertainty-aware analysis of population genetic embeddings.

3 Domain Tasks

Through collaboration with domain experts in population genetics, we identify four primary domain tasks that motivate the design of *SNPScape*.

T1: Identifying feature-driven structure. Analysts seek to understand how specific subsets of genetic variants influence the structure observed in DR embeddings. Existing workflows typically analyze embeddings constructed from the full dataset, providing limited insight into which features drive clustering patterns.

T2: Evaluating embedding robustness and individual stability. Analysts aim to assess whether observed group structure and individual placements remain stable under perturbations such as SNP subsampling. Stability information can be used as a quality filter, identifying individuals whose positions cannot be reliably interpreted in a single static embedding, and which may be candidates for exclusion or further investigation in downstream analysis.

T3: Detecting and characterizing admixture. Analysts seek to visualize patterns in populations with known histories of admixture, such as American populations with components of Native American, European, and African ancestry, or European populations with potential African legacy. The goal is to trace sources of admixture through variant-subset projections and to observe how admixed populations behave differently from other populations under population-specific variant filters.

T4: Assessing population overlap. Analysts aim to communicate the extent to which populations overlap in genetic space and to avoid interpretations of clusters as discrete or well-separated. Standard DR plots often visually emphasize separation, making it difficult to convey continuity and shared variation across populations.

4 Methodology

To support our domain tasks (T1–T4), we introduce a multi-view interactive dashboard that links three complementary visualizations: variant overlap Euler diagrams, dimensionality reduction embeddings, and embedding stability overlays.

4.1 Dataset and Preprocessing

We use genotype data from the 1000 Genomes Project (1KGP) [8] as the primary application context. SNPs are filtered to common variants using a minor allele frequency threshold ($\text{MAF} \geq 0.05$), at the global, continental, and subcontinental level. Set-based overlap operations determine unique and intersecting SNP subsets at continental and subcontinental granularity.

4.2 Euler Diagram Construction

Our Euler diagram (Fig. 2a) construction pipeline builds on prior work by Kitchens and Coop [19], which introduces a GeoVar-based framework for summarizing population-level variant distributions and visualizing shared and population-specific common variants [2]. For each variant, minor allele frequencies across populations are discretized into five bins using thresholds at 0, 0.01, 0.05, 0.1, and 1.0. This produces a discrete GeoVar code for each variant that encodes its frequency category across populations. The frequency of each GeoVar code is then aggregated to obtain a compact representation of variant distributions.

To construct Euler diagram region sizes, we expand the GeoVar codes into per-population bin assignments. Population combinations are enumerated, and for each combination, we compute the number of variants that satisfy our common-variant threshold of 0.05. These shared counts are converted into region sizes by subtracting contributions from higher-order supersets, ensuring that each variant contributes to exactly one region. This computation is implemented using vectorized operations and row-wise filtering over the expanded GeoVar table. The resulting region counts are provided to the `eulerr` package to fit ellipse-based Euler diagrams [20]. We extract ellipse parameters (center coordinates, radii, and rotation), as well as stress and diagonal error. Population metadata, including population labels and sample sizes, are additionally incorporated in a final JSON export.

Euler diagrams are rendered interactively in the browser using `D3.js` [3]. Ellipse parameters returned by `eulerr` are mapped into screen space by computing global bounds across all ellipses, applying a uniform scaling factor, and translating the layout to center the diagram within a fixed viewport. Each ellipse is rendered as an SVG path using an elliptical arc command, with rotation applied via SVG

transforms. Visual encodings, including stroke color, width, and dash patterns, are derived from population attributes. Interactive behaviors are implemented using D3 event handlers: hovering highlights individual populations and displays a tooltip containing metadata such as population description, sample size, and counts of shared and unshared common variants; mouse movement updates tooltip position; and clicking selects a population and triggers linked interactions with other coordinated views in the system. This design enables direct exploration of population overlap while supporting integration with downstream analyses.

4.3 DR Generation for SNP Subsets

To analyze how variant selection influences population structure, we generate DR embeddings from population-specific SNP subsets as well as matched random subsets of equal size. Genotype matrices are constructed for each population by aggregating biallelic SNPs across chromosomes into an $n \times p$ matrix, where n is the number of individuals and p is the number of selected SNPs.

We compute two types of DR embeddings: PCA and UMAP. To accommodate the high dimensionality of the genotype matrices, PCA is implemented using `scikit-learn`'s `IncrementalPCA` [28], which processes data in mini-batches to reduce memory usage. Specifically, the model is fit iteratively using a batch size of 16, followed by a second pass to transform the data into a two-dimensional embedding. UMAP embeddings are generated using the `umap-learn` library with parameters $n\_components = 2$, $n\_neighbors = 15$, $min\_dist = 0.1$, and a fixed random seed for reproducibility.

To isolate the effect of SNP subset selection, we also construct random SNP subsets matched in size to population-specific unique variant sets (Fig. 3). Random subsets are generated by uniformly sampling genomic positions from the full set of biallelic SNP positions across all chromosomes. For each subset size, we generate five replicates to account for stochastic variability. These sampled SNP sets are used to construct corresponding genotype matrices and compute DR embeddings using the same pipeline described above.

The resulting PCA and UMAP embeddings are visualized as two-dimensional scatterplots, where each point represents an individual and coordinates correspond to the first two principal components or the two UMAP dimensions. For interactive exploration, we use `plotly.express` to create scatterplots colored by population with hoverable metadata, standardized marker sizes, and equal axis scaling. To compare embeddings derived from biologically meaningful SNP subsets with size-matched random subsets, we generate animated scatterplots in which each frame corresponds to a different subset or replicate.

To enable comparison between embeddings, we apply Procrustes alignment [13] to map each subset-derived embedding onto a common reference embedding (e.g., the full SNP set or a baseline subset). This alignment removes differences due to rotation, translation, and scaling, ensuring that observed changes in point positions reflect differences in SNP composition rather than arbitrary embedding transformations. This step facilitates a meaningful visual comparison of structure and individual placement across subset conditions.

This design enables controlled comparisons between embeddings derived from biologically meaningful SNP subsets (e.g., population-specific variants) and size-matched random subsets. By holding the number of variants constant, differences in embedding structure can be attributed to the composition of SNPs rather than their quantity, supporting analysis of feature-driven variability in population structure.

4.4 Stability Estimation Under SNP Subsampling

For stability plots (Fig. 1a), we estimate embedding stability via repeated SNP subsampling and alignment:

1. Draw B random SNP subsets $S_b \subseteq S$ at a fixed sampling rate.
2. Recompute DR embedding coordinates for each subset.
3. Align embeddings with a population-wise Procrustes transformation.
4. Compute per-individual positional variability across aligned runs.

For individual i with aligned 2D positions $\mathbf{x}_i^{(1)}, \ldots, \mathbf{x}_i^{(B)}$, we first compute the centroid:

$$\bar{\mathbf{x}}_i = \frac{1}{B} \sum_{b=1}^{B} \mathbf{x}_i^{(b)}.$$

We then define the unnormalized positional variance as:

$$\mathrm{Var}(i) = \frac{1}{B} \sum_{b=1}^{B} \left\| \mathbf{x}_i^{(b)} - \bar{\mathbf{x}}_i \right\|_2^2.$$

To account for differences in overall embedding scale across populations, we normalize this variance by a population-level scale factor. For a population p with embedding coordinates $\mathbf{X}_p$, we define:

$$\mathrm{Scale}(p) = \mathrm{tr}\big(\mathrm{Cov}(\mathbf{X}_p)\big),$$

where $\mathrm{Cov}(\cdot)$ denotes the sample covariance matrix and $\mathrm{tr}(\cdot)$ is the matrix trace.

The normalized stability measure for individual i is then:

$$\mathrm{NormSpread}(i) = \frac{\mathrm{Var}(i)}{\mathrm{Scale}(p)}.$$

Higher values of $\mathrm{NormSpread}(i)$ indicate less stable placement and greater representational ambiguity in the 2D embedding. For interpretability and visualization, individuals are categorized as *Stable* or *Unstable* based on a threshold applied to this normalized spread.

4.4.1 Alignment considerations for UMAP

Since UMAP is stochastic and locally preserving, global Procrustes alignment can be unstable, especially when inter-cluster distances are weakly meaningful across runs. We therefore use group-wise Procrustes alignment anchored by superpopulation partitions.

4.5 Contour-based Visualization of Subsampling Variability

To visualize how embedding structure varies across repeated SNP subsamples, we overlay density contours on aligned PCA and UMAP embeddings (Fig. 2b and 2c). After Procrustes alignment, all subsampled embeddings are plotted as a low-opacity background scatterplot and superpopulation-level spatial density is estimated using kernel density estimation (KDE) [27] separately for each continent. Contours are computed over the aligned coordinates (x_1, x_2) using `seaborn`'s `kdeplot`.

For each superpopulation, we draw multiple contour levels over the embedding to summarize the distribution of points across subsampling runs. These contours provide a compact visual summary of where individuals from a given group tend to appear and how much that placement varies across iterations. Broader or more diffuse contours indicate greater positional variability, whereas tighter contours indicate more stable embeddings. In addition to contours estimated from all aligned points, we also generate centroid-based contour plots by first computing the centroid of each superpopulation within each subsample and then applying KDE to the collection of centroids across runs. This centroid-based view suppresses within-group point-level noise and emphasizes variation in group-level placement across subsamples. Together, these contour views complement per-individual stability scores by providing a group-level visual summary of embedding uncertainty under SNP subsampling.

5 System Design: The *SNPScape* Dashboard

We present *SNPScape*, an interactive visualization dashboard for exploring population genetic structure, variant overlap, and embedding stability in the 1000 Genomes Project. The system is organized into three arranged views: (1) the Global Common Variant View, (2) the Superpopulation Explorer, and (3) the Continent-Level Explorer. Together, these views support a multi-level analysis workflow that moves from global structure to increasingly fine-grained population comparisons.

5.1 Global Common Variant View

The global view provides an overview of population structure using common variants shared across all populations. This view presents PCA and UMAP embeddings computed from globally shared SNPs, allowing users to examine large-scale structure and relationships between major population groups, supporting T2. Stability-aware overlays (Fig. 1a) can be applied to these embeddings to identify *Stable* or *Unstable* points, visualizing variability in individual placement across SNP subsamples, supporting T3. These plots are accompanied by a slider to set the stability threshold value. To support interpretation of variability at the group level, the view includes optional contour overlays that summarize the spatial distribution of each superpopulation (Fig. 1b). These contours can be displayed either at the individual level (aggregating all aligned points) or at the centroid level (aggregating per-subsample population centroids) to provide a visual embedding for continental variability across subsampling runs. To account for Procrustes alignment considerations of UMAP, the UMAP panel supports additional group-wise aligned contour plots (Fig. 1c). These embeddings allow users to assess and compare distributions across subsample runs for each continent group with less influence from stochasticity factors.

This view supports high-level tasks such as identifying broad population clusters, assessing global separability, and establishing a baseline for comparison with subset-specific embeddings in downstream analyses. By combining point-based embeddings with stability overlays and contour summaries, the global view enables users to assess both the overall structure and the robustness of population relationships.

5.2 Superpopulation Explorer

The superpopulation explorer enables analysis of variant overlap and embedding structure at the continental level (Fig. 2). The 1KGP continents consist of American (AMR), African (AFR), East Asian (EAS), European (EUR), and South Asian (SAS) populations. This view supports T1 by linking variant overlap in Euler diagrams (Fig. 2a) to embedding structure in SNP subset DR plots (Fig. 2b and 2c). The Euler diagram summarizes the number of common variants unique to each superpopulation as well as those shared across combinations of superpopulations, contributing to T4.

Users can interactively select a superpopulation from the Euler diagram to view corresponding DR plots of variant subsets unique to a continental group (e.g., variants common only in AFR). The linked DR plots enable direct inspection of how variant composition influences observed population structure, allowing for T3 analysis.

5.3 Continent-Level Explorer

The continent-level explorer further supports T1, T3, and T4 by enabling finer-grained analysis within superpopulations, allowing users to examine how subpopulation-specific SNP subsets influence structure at a more detailed scale. For a selected superpopulation, this view presents Euler diagrams of subpopulation-level variant overlap alongside DR embeddings computed from subpopulation-specific SNP subsets (Fig. 3b). This enables users to explore heterogeneity within continents and to identify patterns that may not be visible at the global or superpopulation level.

6 Evaluation

We conduct an expert-based evaluation to assess whether *SNPScape* supports interpretation of feature-driven population structure and embedding stability beyond traditional DR visualizations. The study compares two conditions: static PCA and UMAP plots and the interactive *SNPScape* dashboard.

6.1 Study Design and Procedure

The goal of the study is to understand how domain experts use the system to complete analysis tasks and whether the system improves insight generation [31] and the interpretation of population genetic structure. Each session begins with a walkthrough of the dashboard's three coordinated views: the Global Common Variant View, the Superpopulation Explorer, and the Continent-Level Explorer, including the stability marker slider, contour display options, and population toggle features. Participants are asked to complete self-defined analysis tasks using both the traditional baseline and the interactive dashboard, narrating their reasoning aloud.

For each task, participants perform the analysis under two conditions: (1) static PCA/UMAP plots (Plotly-rendered, interactive for zoom and hover but without SNP-subset selection, stability overlays, or contour features) and (2) the interactive SNPScape dashboard. The order of conditions is varied across tasks to mitigate learning and ordering effects. Participants are also asked to describe alternative baseline tools and methods that could not be prepared in advance. At the end of each task, participants provide qualitative feedback and answer questions regarding subjective workload through the NASA Task Load Index (NASA-TLX) [16] and software usability through the System Usability Scale (SUS) [4] for the three conditions described. Each session is approximately 45–60 minutes.

6.2 User-Defined Tasks

For each user study session, we ask participants to describe the analysis tasks or research questions they would like to investigate using the dashboard and the baseline. In the following, we summarize the primary tasks identified.

6.2.1 Outlier detection and unusual individual identification

A common first step when examining the 1KGP dataset is to assess whether any individuals or subpopulations display unexpected behavior in the embedding, such as unusual clusters, isolated points, or individuals that appear to have been placed far from their labeled group. This kind of exploratory scan is typically performed using static PCA or UMAP, where users visually inspect the plot for anomalies. However, generating these views, identifying the specific individuals responsible for unusual features, and connecting them across PCA and UMAP plots requires substantial manual effort.

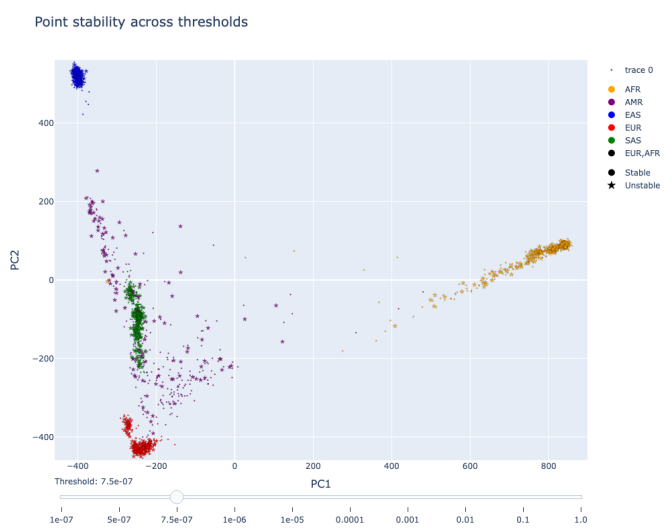
6.2.2 Exploring uncertainty in PCA embeddings

A second recurring question is whether individual positions in a PCA embedding remain stable if the SNP set used to compute the embedding is perturbed. This question has practical implications for downstream analysis: individuals with highly variable positions under subsampling may represent ambiguous or unreliable data points. While stability analysis has been explored analytically [14], it is not a standard component of population genetic visualization workflows. This task is identified by multiple participants as a key concern, particularly in relation to the broader interest of uncertainty quantification in DR and the potential to extend to challenges of positional reliability in settings such as ancient DNA with partial SNP overlap.

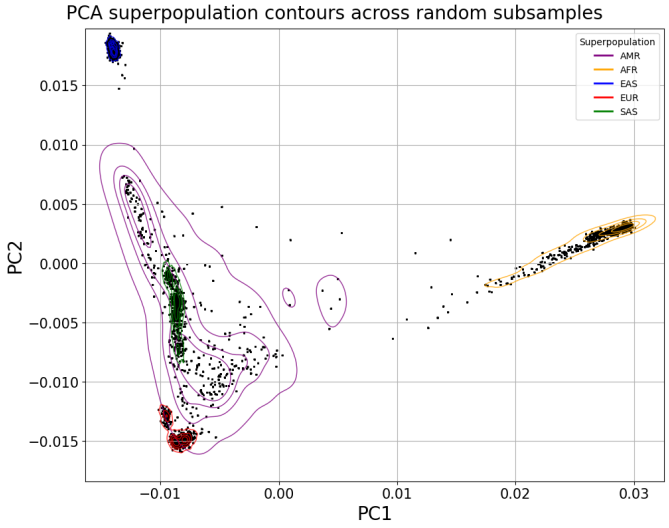
6.2.3 Admixture and historical gene flow analysis

Multiple participants independently define tasks centered on characterizing admixture patterns using population-specific SNP subsets. American populations in the 1KGP dataset carry varying proportions of Native American, European, and African ancestry due to historical admixture, making them a natural test case for feature-driven structure analysis. A central question is whether projecting individuals onto embeddings derived from AMR-specific common variants, or from subpopulation-specific variants within the Americas, reveals admixture gradients and population signals that are not visible in the global embedding.

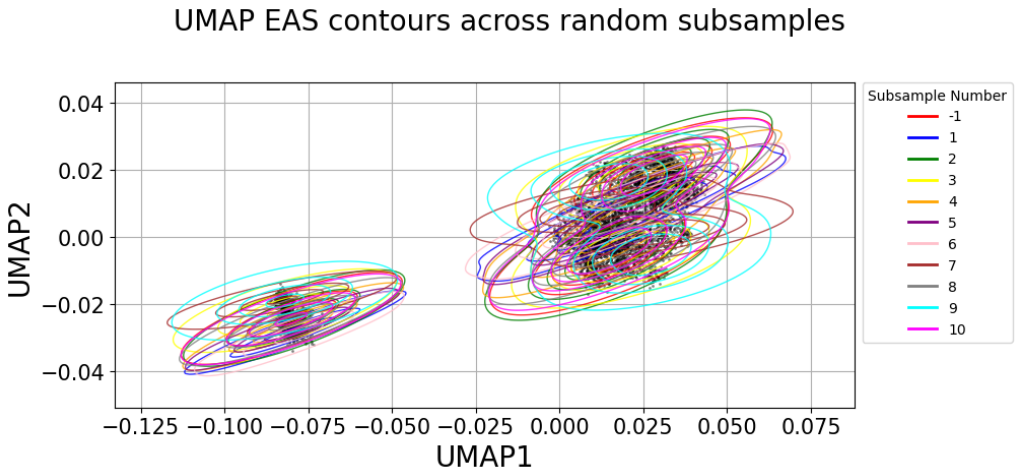
Beyond American populations, participants also express interest in using this approach to investigate more subtle historical gene flow signals in other regions. For example, detecting African genetic contributions to Iberian populations requires identifying weak, population-specific signals that may not be visible in embeddings derived from globally shared variants. Such questions are typically addressed using formal statistical tools, and participants are interested in whether variant-specific embeddings can provide a complementary, visualization-first exploratory perspective.



(a) PCA individual stability markers.



(b) PCA superpopulation contours across random SNP subsamples.



(c) UMAP EAS contours across random SNP subsamples.

Fig. 1: Global uncertainty views in SNPScape. (a) Stability markers identify individuals with high positional variability across subsampled PCA embeddings relative to the user selected stability threshold . (b) Superpopulation-level contours summarize the distribution of points in aligned PCA embeddings across random SNP subsamples. (c) UMAP contour overlays show variability across group-wise Procrustes aligned SNP subsamples for the EAS superpopulation, highlighting differences in embedding structure across runs.

6.2.4 Variation and structure within admixed populations

Another task focuses on understanding how admixed individuals behave differently from other populations when the analysis is restricted to continent-specific variant filters. The motivating observation is that admixed individuals are often described as falling intermediate between reference populations in a global PCA, but this characterization depends on using global common variants as the reference frame. When the analysis is instead conducted using continent-specific common variants, admixed individuals may no longer appear intermediate, revealing a different view of their genetic composition. There is an interest in whether the dashboard can make this shift in perspective visible and interpretable, particularly for understanding how African Americans and other admixed groups appear when viewed through an Africa-specific variant lens rather than a global one.

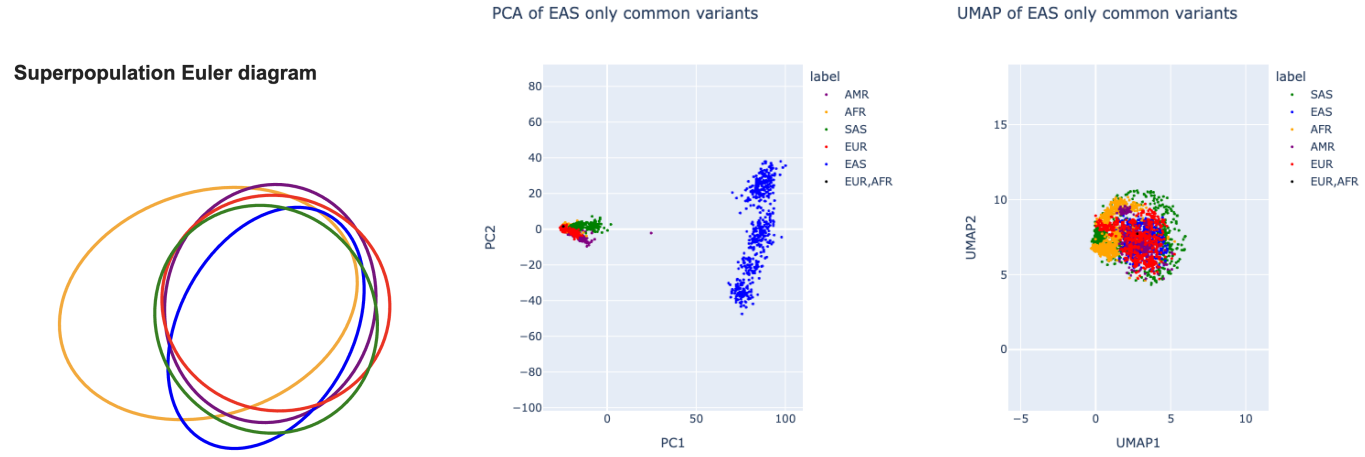
7 Results and Discussion

Overall, we find that SNPScape improves users' ability to generate insights, interpret feature-driven structure, and understand embedding uncertainty compared to traditional baselines. Across all three participants, the dashboard enables more diverse and deeper insights, while the static baseline primarily supports confirmatory observations grounded in prior domain knowledge. The alternative baseline condition further suggested that many analyses performed with SNPScape would either not have been attempted or would have required extensive scripted workflows to approximate.

7.1 Qualitative Feedback

All participants provided positive overall assessments of the system and highlighted several recurring themes regarding its impact on analysis workflows. A central benefit was the ability to perform tightly coupled, side-by-side comparisons across DR methods, variant subsets, and population labelings. This capability enabled faster cross-view reasoning that would otherwise require substantial manual coordination across tools or displays, allowing participants to more efficiently interpret differences between embeddings.

Participants also emphasized the value of the stability and uncertainty visualizations, which made properties of the underlying DR methods more directly observable. These views supported intuitive reasoning about embedding robustness and variability, enabling participants to distinguish between stable structure and representations sensitive to perturbation. In particular, the ability to visualize the distribution of



(a) Superpopulation Euler diagram summarizing shared and population-specific SNP sets.

(b) PCA embedding using EAS-specific common variants.

(c) UMAP embedding using EAS-specific common variants.

Fig. 2: The Superpopulation view links variant overlap with embedding structure. (a) The Euler diagram summarizes shared and population-specific SNP sets across superpopulations. (b) PCA embedding derived from EAS-specific common variants reveals feature-driven separation, with EAS individuals forming a distinct structure relative to other populations. (c) UMAP embedding of the same SNP subset emphasizes local structure and shows reduced global separation.

individuals across subsampled embeddings through contours provided a novel perspective on population structure.

Beyond supporting specific analytical tasks, the system prompted broader reflection on analytical practices and assumptions. Participants reported reconsidering how DR outputs are interpreted and expressed interest in incorporating stability-aware workflows, such as systematic subsampling, into their analyses. Participants with extensive prior experience with the dataset were able to identify new patterns and lines of inquiry, suggesting that the system supports continued discovery and deeper engagement with familiar data. Together, these findings indicate that SNPScape not only enables users to examine the data from alternative perspectives but also expands the scope of analyses that participants are able to perform.

Participants also provided several concrete suggestions for extending the system. These included enhancing selection capabilities in the Global Common Variant View to support subpopulation-level color schemes, and introducing brushing and linking of individuals so that interactions in one embedding (e.g., the left PCA panel) highlight corresponding points in another (e.g., the right UMAP panel). There was also interest in expanding the range of variant subset operations, particularly to include shared or intersecting SNP sets between populations rather than only population-specific variants. Finally, participants noted that extending the system to additional datasets would further broaden its applicability.

7.2 Quantitative Evaluation

7.2.1 NASA-TLX

We administered the NASA-TLX questionnaire to assess perceived workload across conditions (Fig. 4). Given the small sample size (6 scores per question and condition), Friedman tests and Bonferroni-corrected Wilcoxon signed-rank post-hoc comparisons are reported as indicative rather than conclusive. In cases where no within-subject differences were observed for Wilcoxon signed-rank tests, p-values were conservatively set to 1.0.

Based on average scores across questions and conditions, SNPScape is associated with lower frustration, temporal demand, and physical demand, suggesting that tasks can be completed more easily and efficiently once users are oriented to the interface. In contrast, SNPScape shows higher mental demand and effort, which is consistent with participants engaging more actively with the system's interactive features and exploring more complex analytical questions. SNPScape additionally achieved a higher average performance score, indicating that the tool was more successful in accomplishing the user-defined tasks compared to the static and alternative baselines.

Friedman tests indicated significant differences for five of six dimensions, with fairly strong agreement across participants (Kendall's $W = 0.75$–0.78) (Table 1). Effects were primarily driven by the contrast between the alternative baseline and the two interactive conditions, but consistent directional trends suggest that SNPScape shifts workload from mechanical and time-intensive effort toward more engaged analysis.

7.2.2 SUS

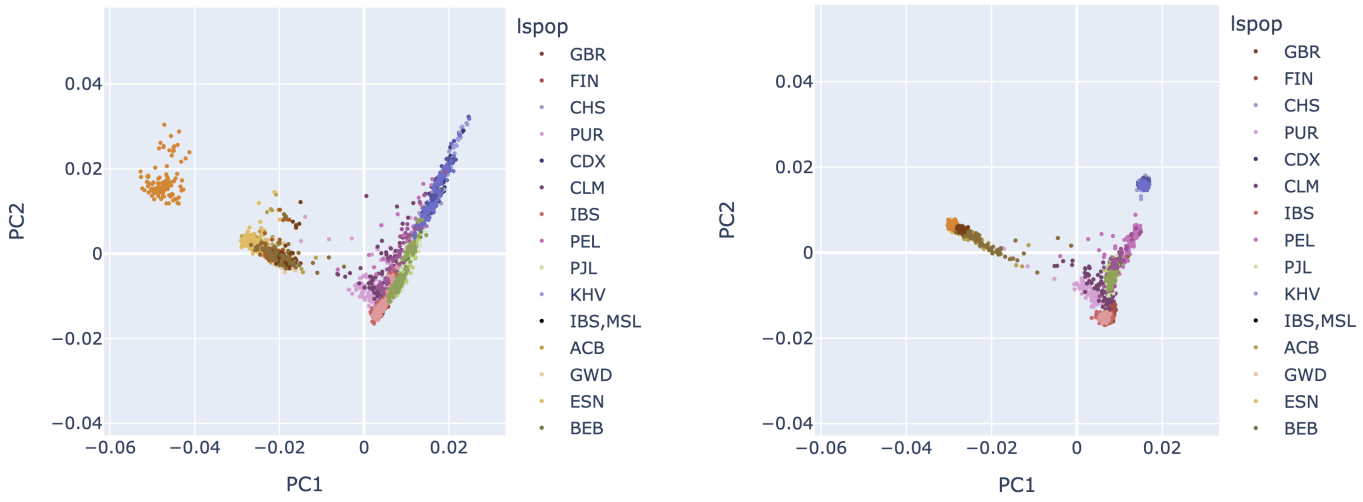
System Usability Scale scores were computed using the standard 0–100 scoring procedure (Fig. 5). SNPScape achieved the highest results, with a mean score of 83.8 (SD = 7.2), while the static baseline achieved 71.7 (SD = 9.7). The alternative baseline scored substantially lower (mean = 37.9, SD = 15.8).

Item-level SUS responses (Fig. 6) indicate that SNPScape was rated higher on usability dimensions including use frequency, ease of use, integration, consistency, learnability, and perceived lightweight interaction. These patterns indicate that participants found SNPScape not only more useful but also better integrated and more efficient to work with. The static baseline scored slightly higher on simplicity and no technical help need, reflecting participants' familiarity with conventional DR methods. The alternative baseline received substantially lower ratings across nearly all items, particularly in ease of use, simplicity, and lightweight interaction, highlighting the perceived burden of traditional multi-tool analytical workflows.

Friedman tests reveal significant condition effects for eight of ten SUS items, again primarily driven by the contrast with the alternative baseline (Table 2). However, the consistent improvements in mean scores suggest that SNPScape provides a more usable and effective interface than standard static visualizations, with the primary trade-off being an increase in initial learning effort.

7.3 Insights

For all three user study sessions, participants narrated their thought processes aloud, enabling identification and counting of distinct insights generated under each condition. As summarized in Table 3, participants consistently generated substantially more insights when using SNPScape compared to the static baseline.



(a) PCA embedding using YRI-specific common variants.

(b) PCA embedding using a size-matched random SNP subset.

Fig. 3: Continent-level comparison of subpopulation-specific and size-matched random SNP subsets. (a) The YRI-specific common variant embedding reveals fine-grained subpopulation structure that differs from global projections. (b) A size-matched random SNP subset produces a different structure, suggesting that the pattern in (a) may be driven by SNP composition rather than subset size alone.

Table 1: NASA-TLX workload ratings (1–10 scale) across conditions: static baseline (B), SNPScape (SS), and alternative baseline (B2). Values report mean scores per dimension, where lower scores indicate lower perceived workload, except for Performance. Friedman tests (χ^2) assess overall condition effects, with Kendall's W indicating agreement across participants. Post-hoc Wilcoxon signed-rank tests with Bonferroni correction compare pairs of conditions. Significant differences are primarily driven by the alternative baseline, which exhibits substantially higher workload across most dimensions.

Item	Mean B	Mean SS	Mean B2	χ^2	p	W	SS–B p	SS–B2 p	B–B2 p
Mental	3.67	5.17	8.50	8.43	0.015	0.778	1.000	0.188	0.094
Physical	3.33	2.00	6.33	6.50	0.039	0.778	1.000	0.188	0.750
Temporal	3.00	2.50	8.33	8.67	0.013	0.778	1.000	0.188	0.094
Performance	4.33	6.33	4.50	4.57	0.102	0.361	0.188	0.938	1.000
Effort	3.67	4.67	8.67	9.33	0.009	0.778	1.000	0.094	0.094
Frustration	3.17	2.00	6.50	10.18	0.006	0.750	1.000	0.094	0.094

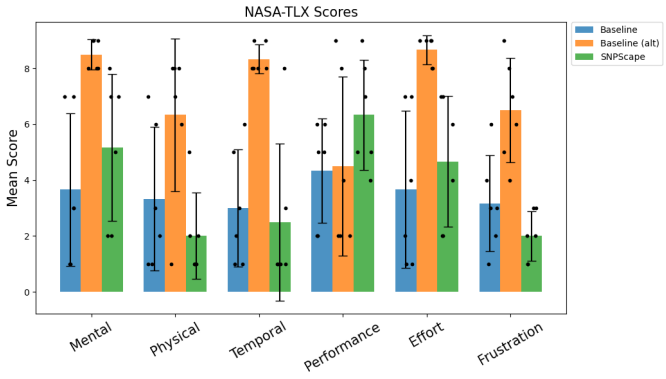


Fig. 4: Average NASA-TLX scores per condition. Error bars indicate standard deviation. Points represent each of the individual scores. Lower scores indicate less demand, except for Performance.

These differences reflect not only an increase in the number of observations, but also a shift in the types of insights participants were able to generate. Under the static baseline, insights were primarily limited to broad, confirmatory interpretations of global structure. In contrast, SNPScape enabled more specific, data-driven, and multi-level reasoning.

We organize these findings into four recurring analytical patterns observed across participants and tasks:

7.3.1 SNPScape enables identification and interpretation of outliers through coordinated spatial and stability plots.

Participants engaging in exploratory outlier detection found that the static baseline supported broad spatial reasoning about population distributions. With SNPScape, participants identified specific outlying individuals and interpreted their positions with additional context from uncertainty embeddings. Stability markers supported outlier interpretation by distinguishing stable from unstable embeddings. Participants observed that individuals in small, isolated clusters or boundary positions tended to show higher positional variance under subsampling, providing additional evidence that these placements reflect a level of representational ambiguity. The side-by-side comparison of PCA and UMAP plots additionally allowed participants to compare individuals across the two DR methods, identifying whether unusual patterns were consistent across embeddings. This connection between spatial outliers and stability score was not available under the baseline condition.

7.3.2 Uncertainty visualizations reveal both population-level and individual-level variability in embeddings.

All three participants engaged with the stability and contour features to explore embedding robustness. Under the static baseline, participants did not have access to direct measures of embedding stability,

Table 2: SUS item-level results (1–5 scale) across conditions: static baseline (B), SNPScape (SS), and alternative baseline (B2). Values report mean ratings per item, with higher values indicating more positive usability perceptions. Friedman tests (χ^2) evaluate overall condition effects, with Kendall's W measuring agreement across participants. Post-hoc Wilcoxon signed-rank tests with Bonferroni correction assess pairwise differences. SNPScape and the static baseline show consistently higher usability ratings than the imaginary baseline, with differences between the two interactive conditions remaining smaller and not statistically significant after correction.

Item	Mean B	Mean SS	Mean B2	χ^2	p	W	SS–B p	SS–B2 p	B–B2 p
Use frequency	4.00	4.67	2.17	9.58	0.008	0.583	0.375	0.188	0.188
Simplicity	4.33	4.17	2.17	9.48	0.009	0.750	1.000	0.094	0.094
Ease of use	4.00	4.50	1.83	9.24	0.010	0.694	1.000	0.094	0.188
No technical help needed	4.67	4.33	3.83	4.00	0.135	0.028	1.000	1.000	0.375
Integration	2.83	4.17	2.17	6.74	0.034	0.333	0.562	0.188	0.750
Consistency	4.00	4.67	3.17	8.10	0.017	0.528	0.938	0.094	0.750
Learnability	3.33	4.17	2.50	4.10	0.129	0.194	0.938	0.375	1.000
Lightweight	3.67	4.33	2.00	10.38	0.006	0.861	1.000	0.094	0.094
Confidence	4.17	4.17	2.83	6.78	0.034	0.194	1.000	0.375	0.188
No prior learning	3.67	4.17	2.50	7.90	0.019	0.444	1.000	0.188	0.188

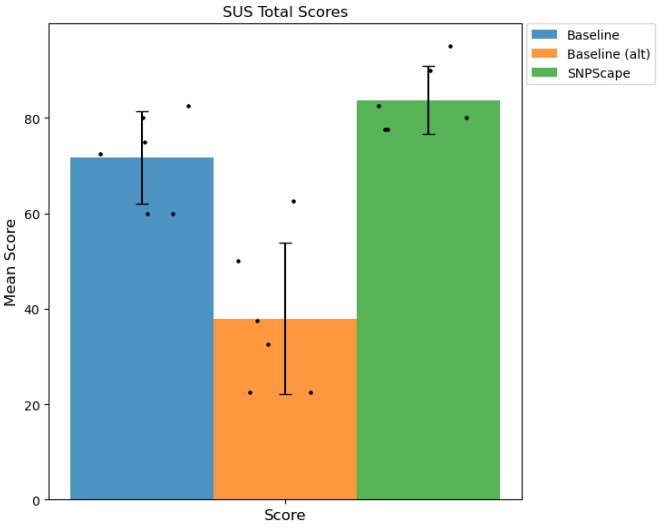


Fig. 5: Average SUS scores per condition. Scores are out of 100. Error bars indicate standard deviation. Points represent each of the individual scores.

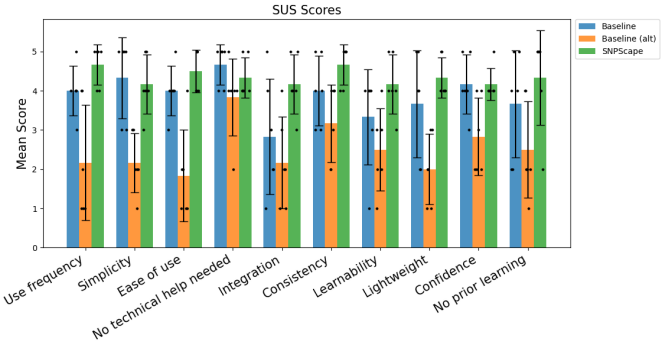


Fig. 6: Average SUS item scores per condition. Error bars indicate standard deviation. Red lines indicate median. Even-numbered questions are inverted for interpretability, making higher scores more desirable across all items.

Table 3: Comparison of insight counts per participant between the static baseline and SNPScape conditions. Differences between SNPScape (S) and baseline (B) indicate consistent increased insight generation for the SNPScape condition.

Participant	**Baseline**	**SNPScape**	Δ (S – B)
P1	**6**	**18**	**12**
P2	**5**	**18**	**13**
P3	**8**	**15**	**7**

and standard practice PCA outputs were generally treated as fixed representations.

SNPScape produced several consistent observations across participants. First, admixed populations exhibited lower stability in UMAP embeddings compared to non-admixed populations in UMAP embeddings, particularly those with complex or recent admixture histories. Second, instability was found to be individual-specific rather than purely group-level, with some individuals in populations not known for admixture exhibiting substantial positional variation. Participants framed this as information that could serve as a quality filter for downstream analysis.

Contour visualizations provided a group-level complement to the individual stability markers and revealed clear methodological differences: PCA embeddings were compact and stable across subsamples, while UMAP embeddings were diffuse and variable. Participants observed that although UMAP cluster positions shift across runs, the internal cluster structure remains more consistent, providing important context for interpreting intra- versus inter-cluster relationships.

Future work could explore alternative uncertainty measures and encode multiple metrics into a single embedding. Additionally, understanding how different user populations interpret uncertainty visualizations of population genetic data, and how these representations influence decision-making, remains an important research question.

7.3.3 SNP subset-driven embeddings reveal admixture gradients not observable in global projections.

Admixture analysis in American populations emerged as a recurring task. Under the static baseline, participants identified general admixture patterns. Specifically, American individuals are broadly distributed in the global PCA, with some near European clusters, others shifted toward African clusters, and others in between, reflecting varying proportions of Native American, European, and African ancestry. However, participants additionally noted that this interpretation relied heavily on prior knowledge, and they could not attribute signals to specific variant subsets.

Using SNPScape, participants explored embeddings derived from variant subsets that revealed admixture gradients not visible in global projections. AMR-specific SNP subsets exposed clear gradients along the principal components, with individual subpopulations occupying distinct positions along this gradient consistent with their known admixture histories. Comparing the AMR-specific PCA to a size-matched random SNP subset confirmed that the observed gradient is driven by the biological composition of the variant set rather than subset size

alone. Further investigation into subpopulation-specific variant sets revealed that SNPs unique to a given American subpopulation placed most individuals from that group closer to specific non-American reference populations, providing a visualization-first way to trace admixture sources that could serve as a hypothesis-generating step prior to formal quantitative analysis. UMAP additionally revealed subpopulation clustering patterns within the Americas that PCA did not make obvious, illustrating how the two methods expose complementary aspects of the same underlying data.

7.3.4 Continent-specific variant filtering exposes subpopulation structure and challenges global interpretations of population separation.

Participants observed that filtering embeddings to reflect continent-specific variants revealed subpopulation structure that is missed in global embeddings. Under the static baseline, within-continent variation was not as attended to, as global PCA and UMAP plots direct visual attention toward the dominant between-continent structure. Additionally, under the alternative baseline condition, no participant had a workflow to systematically examine how the structure within the continent varies as a function of the variant subset.

With SNPScape, filtering to a superpopulation's unique common variants produced a completely different cluster configuration, including cases where the canonical global PCA triangle disappeared entirely. This finding prompted participants to reflect on how the field routinely treats DR visualizations as definitive rather than conditional on feature selection. Populations that appear as tight, homogeneous clusters in the global view were found to contain their own internal subpopulation structure when viewed through continent-specific variant lenses, supporting a more continuous and context-dependent interpretation of genetic variation.

One participant's task of detecting historical gene flow between African and European populations through African-specific SNP subsets also revealed a current design boundary of the system: the Continent-level explorer exposes embeddings derived from population-unique variants, but detecting subtle historical admixture requires examining variants shared between source and recipient populations. However, participants noted that the visualization exploration served as a useful and confirmatory step, and that extending the system to support shared-variant embeddings would allow for additional historical gene flow questions to be explored through visual analysis.

7.4 Scalability and Computational Constraints

A practical limitation of our system is the computational cost of generating DR embeddings across multiple SNP subsets and subsampling iterations. Due to the high dimensionality of genomic data, embeddings are precomputed offline rather than generated dynamically. While this enables interactive exploration, it limits the flexibility of user-defined subset selection.

One direction we seek to explore is to incorporate preprocessing techniques such as linkage disequilibrium (LD) pruning [29, 32] to reduce redundancy in SNPs and enable more efficient computation. This could support more dynamic workflows where users interactively define SNP subsets and compute embeddings in real time.

7.5 Challenges in Aligning Stochastic Embeddings

Aligning embeddings across subsampling runs is critical for stability analysis, but remains fundamentally challenging for non-linear dimensionality reduction methods such as UMAP. Unlike PCA, which is globally preserving, methods such as UMAP and t-SNE are inherently stochastic and emphasize local structure. As a result, embeddings may differ across runs not only due to feature perturbations (e.g., SNP subsampling), but also due to randomness in the optimization process and sensitivity to initialization. Prior work on stability analysis of stochastic embeddings has shown that such variability can lead to substantially different geometric configurations even when global structure is preserved, complicating direct comparison across runs [15].

A common approach to comparing embeddings is Procrustes alignment, which removes differences in translation, rotation, and scaling. However, this assumes that embeddings differ only by global transformations. For locally preserving methods, this assumption is often violated, as cluster shapes, relative distances, and local neighborhoods may shift in ways that cannot be captured by a single global transformation. Consequently, global Procrustes alignment may obscure meaningful variability or fail to align corresponding structures across embeddings.

In this work, we mitigate this issue by performing group-wise Procrustes alignment using superpopulation labels, allowing each group to be aligned independently. While this improves correspondence at the population level, it introduces additional assumptions. Specifically, it requires predefined or inferred group labels. Furthermore, group-wise alignment does not fully resolve inconsistencies in local geometry within groups. Future work could explore alternative alignment strategies that better respect local structure. Developing methods that jointly account for feature-driven variability and algorithmic instability is an important direction for improving the interpretability of dimensionality reduction in high-dimensional biological data.

7.6 Generalizability to Other Domains

While our system is designed for population genetic data, the underlying framework of linking feature-level summaries, embedding structure, and stability analysis can be applied more broadly to high-dimensional datasets. Many domains rely on DR techniques for exploratory analysis, including single-cell genomics, neuroscience, and machine learning model embeddings [1, 9]. In these contexts, feature selection and stochasticity can similarly influence embedding structure and interpretation.

7.7 Toward More Responsible Interpretation of DR

Overall, our findings suggest that DR visualizations should be interpreted as conditional and context-dependent representations rather than definitive maps of structure. By explicitly linking variant overlap, embedding geometry, and stability, *SNPScape* provides tools for more critical and informed interpretation of population genetic data.

An important open research direction is how to design visualization systems that actively discourage overinterpretation and communicate uncertainty without overwhelming users. Addressing this challenge will be essential for ensuring responsible use of DR methods in genomics and other data-intensive fields.

8 Conclusion

We present *SNPScape*, an interactive visualization framework for interpreting dimensionality reduction embeddings in population genetics. Our system integrates three complementary visualizations of 1KGP data, variant overlap via Euler diagrams, embedding structure via PCA and UMAP, and stability-aware visualizations that quantify variability under SNP subsampling, into a unified, multi-view dashboard. This design enables users to directly connect feature-level diversity with embedding-level structure and individual-level uncertainty.

Through expert evaluation, we show that *SNPScape* not only increases the number of insights generated, but also changes the nature of analysis. Compared to static baselines, participants produced substantially more insights and engaged in more specific, multi-step reasoning that links variant selection, embedding structure, and uncertainty. These qualitative results are complemented by quantitative outcomes that demonstrate that *SNPScape* reduces temporal demand and frustration while improving performance and perceived usability compared to baseline conditions. Together, these findings indicate that *SNPScape* improves analytical efficiency and interpretability without compromising perceived task success.

Our results further demonstrate that dimensionality reduction embeddings should be interpreted as conditional and uncertainty-aware representations. By enabling direct comparison across SNP subsets, *SNPScape* reveals that apparent clustering patterns depend on feature selection, while stability visualizations expose both individual and group-level variability that is otherwise invisible in static plots. These capabilities allow users to identify ambiguous or admixed individuals,

distinguish robust structure from unstable representations, and avoid overinterpreting precise point placement.

Beyond population genetics, our work highlights the importance of integrating uncertainty and feature provenance into visualization systems for high-dimensional data. The *SNPScape* framework provides an approach for linking data composition, low-dimensional representations, and variability, and may be applicable to other domains that rely on dimensionality reduction for exploratory analysis.

We hope that this work encourages more critical and responsible use of DR visualizations and motivates future research on uncertainty-aware, multi-view systems for interpreting high-dimensional data.

REFERENCES

[1] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng et al. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44, 2019. 9

[2] A. Biddanda, D. P. Rice, and J. Novembre. Geographic patterns of human allele frequency variation: a variant-centric perspective. *BioRxiv*, pp. 2020–07, 2020. 2

[3] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011. 2

[4] J. Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996. 4

[5] S. Casillas, R. Mulet, P. Villegas-Miron, S. Hervas, E. Sanz, D. Velasco et al. Pophuman: the human population genomics browser. *Nucleic acids research*, 46(D1):D1003–D1010, 2018. 2

[6] T. Chari and L. Pachter. The specious art of single-cell genomics. *PLOS Computational Biology*, 19(8):e1011288, 2023. 1

[7] L. Clarke, S. Fairley, X. Zheng-Bradley, I. Streeter, E. Perry, E. Lowy et al. The international genome sample resource (igsr): A worldwide collection of genome variation incorporating the 1000 genomes project data. *Nucleic acids research*, 45(D1):D854–D859, 2017. 2

[8] . G. P. Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56, 2012. 2

[9] J. P. Cunningham and B. M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500–1509, 2014. 9

[10] A. Diaz-Papkovich, L. Anderson-Trocmé, C. Ben-Eghan, and S. Gravel. Umap reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS genetics*, 15(11):e1008432, 2019. 1

[11] A. Diaz-Papkovich, L. Anderson-Trocmé, and S. Gravel. A review of umap in population genetics. *Journal of Human Genetics*, 66(1):85–91, 2021. 1

[12] E. Elhaik. Principal component analyses (pca)-based findings in population genetic studies are highly biased and must be reevaluated. *Scientific reports*, 12(1):14683, 2022. 1

[13] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 3

[14] J. Haag, A. I. Jordan, and A. Stamatakis. Pandora: a tool to estimate dimensionality reduction stability of genotype data. *Bioinformatics Advances*, 5(1):vbaf040, 2025. 2, 4

[15] R. A. Hamad, E. Järpe, and J. Lundström. Stability analysis of the t-sne algorithm for human activity pattern data. In *2018 IEEE international conference on systems, man, and cybernetics (SMC)*, pp. 1839–1845. IEEE, 2018. 9

[16] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988. 4

[17] J. Hullman. Why authors don't visualize uncertainty. *IEEE transactions on visualization and computer graphics*, 26(1):130–139, 2019. 2

[18] M. Jung, T. Fujiwara, and J. Jo. Ghostumap2: Measuring and analyzing (r, d)-stability of umap. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 2

[19] J. Kitchens and G. Coop. Visualizing human genetic diversity, May 2023. Blog post. 2

[20] J. Larsson et al. eulerr: area-proportional euler and venn diagrams with ellipses. *R package version*, 6(0):2020, 2020. 2

[21] A. Maćkiewicz and W. Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993. 1

[22] J. H. Marcus and J. Novembre. Visualizing the geography of genetic variants. *Bioinformatics*, 33(4):594–595, 2017. 2

[23] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 1

[24] G. McVean. A genealogical interpretation of principal components analysis. *PLoS genetics*, 5(10):e1000686, 2009. 1

[25] J. Novembre and M. Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40(5):646–649, 2008. 1

[26] L. Padilla, M. Kay, and J. Hullman. Uncertainty visualization. *Computational Statistics in Data Science. Wiley*, 2022. 2

[27] E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962. 3

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011. 3

[29] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007. 9

[30] D. Reich, A. L. Price, and N. Patterson. Principal component analysis of genetic data. *Nature genetics*, 40(5):491–492, 2008. 1

[31] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE transactions on visualization and computer graphics*, 11(4):443–456, 2005. 4

[32] M. Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, 2008. 9