A Design Study to Identify Inconsistencies in Kinship Information: the Case of the 1000 Genomes Project

Michaël Aupetit, Ehsan Ullah, Reda Rawi, Halima Bensmail
Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
{maupetit,eullah,rrawi,hbensmail}@qf.org.qa

ABSTRACT

Genome Wide Association Studies (GWAS) examine genetic variants in different individuals to detect variants associated to specific diseases. The 1000 Genomes project is such a collaborative research effort to sequence the genomes of at least 1000 participants of 26 different ethnicities, to establish a detailed summary of human genetic variation. The kinship information is a measure of individuals ancestor relationships within the considered populations. We study the design of kinship data visualizations allowing the experts to discover anomalies in GWAS data. The visual analysis of the 1000 Genomes Project kinship data reveals inconsistencies which call for a deeper analysis of the data quality within this project.

1 INTRODUCTION

In genetic epidemiology, Genome-Wide Association Studies (GWAS) examine many genetic variants in different individuals to see if any variant is associated with a specific trait like a disease. Single nucleotide polymorphism (SNP) genotyping is used to identify such variants.

One of the many challenges in GWAS relies on *representative sampling* of a subset of individuals from a population. Important information like the frequency of alleles in the population are estimated from these representative samples. The problem is that these estimates are subject to inflation or deflation when the SNP data are derived from individuals with unreported or unestimated familial relationships or with admixed ancestry, potentially leading to unwanted population stratification. Therefore it is essential that the samples are composed of individuals with as weak as possible familial relationship or admixed ancestry.

The kinship coefficient measures the degree of ancestry relationship between two individuals [10]. In the GWAS of interest, we expect it to be low for all pairs of individuals in all population samples.

In this paper we design visualizations of the kinship information both within and between population samples of GWAS, in order to enable visual detection of inconsistencies in the population samples that could undermine any further analyses. We use these visualizations to analyze the kinship-focused quality of the samples in the 1000 Genomes Project, an international research effort to establish the most detailed catalog of human genetic variation [12]. We found striking anomalies in the samples that call for a deeper analysis of the kinship processing and the data quality within this project.

2 PROBLEM CHARACTERIZATION: ANALYZING KINSHIP DATA

In GWAS, the expert has to analyze P population samples $S = \{S_1, \ldots, S_P\}$, each sample $s \in S$ being a set of n_s individuals together with their genetic and clinical features. In our design study, each sample is solely characterized by its $n_s \times n_s$ kinship matrix and the identity of the sampled population.

Table 1: **kinship coefficient and lineage dissimilarity**

i, j relationship	kinship coefficient Φ_{ij}	Lineage dissimilarity d_{ij}
Self; Mono-zygote twins	1/2	0
Parent-offspring; Full siblings	1/4	1
Half siblings; Uncle-nephew; Double 1st cousins	1/8	2
Grandchild-grandparent; First cousins	1/16	3
Second cousins	1/64	5
None	0	$+\infty$

A kinship matrix contains similarity measures between 0 and 1 for all pairs of individuals, modeling either close (1) or distant (0) pedigree relatedness [10]. When pedigree records are available, the kinship coefficient between two individuals can be calculated according to the genetic rules postulated by Mendel: for an individual, one copy of each gene is inherited from the father and one from the mother independently and at random. We can assess the probability p_k that two individuals share k genes copies of a common ancestral gene. These genes are said *identical by descent* (IBD). From these probabilities we can compute the kinship coefficient $\Phi_{ij} = \frac{1}{2}p_2 + \frac{1}{4}p_1$ related to individuals i and j which is the probability for two genes to be IBD when genes are selected at a given locus picked at random from each of the two individuals. The higher the kinship coefficient, the closer the individuals in the lineage.

The pedigree-based kinship coefficient between two individuals takes specific values according to their particular familial relationship as shown in the table 1. In a GWAS however, the actual pedigree information relating all the individuals in a sample is usually unknown but the kinship coefficients can be estimated using available genetic information like SNP [13].

To avoid bias in the study of population genomes, it is usual to setup a sampling procedure which selects individuals in the population as genetically different as possible. In that way the observed common genetic characteristics are more likely to be a general trait of the population than a specific trait of the sample. Therefore **the expert expects homogeneity of the samples and that all pairwise kinship coefficients of each population sample should be close to zero**.

Our design problem is to **allow the expert to visually evaluate how much this assumption is valid in the GWAS data at hand** so to decide on a course of action.

3 DATA ABSTRACTION AND VISUAL ENCODING

In this section we describe how we encode the kinship information (3.1) and how we design three visualizations to emphasize kinship anomalies: a between-sample lineage dissimilarities view (3.2); a distribution of lineage dissimilarities view (3.3); and a map of within-sample kinship similarities view (3.4). Kinship data being concerned with relationships between individuals, the node-link diagram representation naturally comes to mind [15]. Node-link

diagrams encode nodes' similarity through presence or absence of links, while intensity of the similarity can be encoded by the color or thickness of the edges between pairs of nodes, or by edge bundles between groups of nodes. In the MizBee synteny browser [9], chromosomes are arranged along concentric rings and connected through edge bundles to encode their relations. Adjacency matrix or heat maps are also standard ways to represent similarity data using color coding of cells in a matrix [14]. However, the position is known to be the most efficient visual encoding [16] for quantitative variables. So we propose to encode values of the quantitative kinship similarity information as points' proximity in a scatterplot using classical Multi-Dimensional Scaling (MDS) [6] (3.2 and 3.4). Although MDS can lead to misleading distorted representations [1] [7] [2], we will see that in our case the distortions are low enough so the MDS plots happen to be trustworthy representations of the kinship similarities. Now we present in details the kinship data encoding and the three static kinship data visualizations that users found efficient and expressive enough to catch striking anomalies in the kinship data.

3.1 kinship information encoding

Each population-sample $s \in S = \{S_1, \ldots, S_P\}$ comes as an $n_s \times n_s$ kinship matrix $\mathbf{M_s}$. Each cell of the kinship matrix $\mathbf{M_s} = \{m_{ij}\}$ is an estimation from genomic data of the actual kinship coefficients $2\Phi_{ij}$ given in the table 1. The real Φ_{ij} values are not available because the actual lineage of each individual in a sample is not known. m_{ij} can take any value between 0 and 1.

As the expert wants to detect similarities or differences between the population-samples, we need to be able to compare population-samples to each other, *i.e.* to visually compare the content of kinship matrices of *different sizes*. Therefore we first transform the kinship matrix $\mathbf{M_s}$ into a *lineage dissimilarity* matrix $\mathbf{D_s} = \{d_{ij}\}$ of values rounded to the nearest integer applying $d_{ij} = \lfloor 0.5 - \log_2(m_{ij}) \rfloor$ to non-zero values m_{ij} and arbitrarily setting $d_{uv} = C$ where $C = 1 + \max_{i,j,m_{ij}\neq 0}(d_{ij})$ when $m_{uv} = 0$. In this way we map the kinship coefficient m_{ij} into a *linear scale* where the lineage dissimilarity value d_{ij} is a more intuitive encoding of a distance between i and j in the ancestor tree as shown in the table 1. The integer-rounding enables *counting* pairs of individuals in a sample which have similar kinship similarity values.

To cope with the different kinship matrix sizes across the population-samples, we either consider a qualitative or a quantitative approach. In the former case, we resort to the visual perception of the graphical representation of each sample (view (3.4)) to accommodate for their different size by driving the visual focus on the shape formed by the items displayed rather than on the exact number of these items. In the latter case (views (3.2) and (3.3)), we generate a vector-based representation of each sample s where we count the frequency of each (rounded) lineage dissimilarity values in $\mathbf{D_s}$. The $d_{ij} \in \mathbf{D_s}$ values lie in the range $[1, C]$ so we define a C-dimension vector v_s representation of s for which the kth coordinate $v_s[k]$ is the proportion of occurrence of the value k in the upper-triangular part of $\mathbf{D_s}$: $\forall k \in \{1, \ldots, C\}, v_s[k] = \frac{2|\{d_{ij} | i<j, d_{i,j}=k\}|}{n_s(n_s-1)}$. Notice that $\sum_{k\in\{1,\ldots,C\}} v_s[k] = 1$.

3.2 Between-sample lineage dissimilarities view

This view is designed to qualitatively compare the population samples by visualizing their pairwise similarities in terms of the within-sample distribution of their lineage dissimilarities.

Geometrical encoding: We use the classical Multi-Dimensional Scaling (MDS) technique [6] as a mean to spatialize the similarities between the population-sample vectors v_s and visualize them as a scatterplot. We compute the Euclidean pairwise distance matrix between the P C-dimension vectors $v_s[1, \ldots, C]$ representing the population samples. We use MDS to visualize these vectors as a scatterplot whose axes are the first and second eigen vectors corresponding to the highest eigenvalues.

Rendering: In this plot, axes are not displayed on purpose to focus the attention of the expert on the distances between the points which encode data similarities. The points are centered in a square frame to avoid aspect ratio bias, and scale is such that if displayed, unit vectors on x and y axes would appear of equal length on the screen. Population codes (Table 2) are displayed near their corresponding points to help link similarity information to external expert knowledge.

Expected interpretation: Exploiting the Gestalt perception law of proximity [16], this qualitative overview representation of the population samples is designed for the expert to visually detect possibly anomalous samples as outlying points or sets of points forming a cluster apart from the core cluster of the other samples.

3.3 Distribution of lineage dissimilarities view

This view is designed to emphasize the presence of many unexpectedly low lineage dissimilarities in a sample and to quantitatively compare the samples based on this information.

Geometrical encoding: This view is a color-coded stacked bar graph showing for each population sample, the proportions $v_s[k]$ of pairwise relationships with the most anomalous rounded lineage dissimilarity values $k = \{1 \ldots 5\}$ according to table 1. Each sample s is represented as an horizontal bar divided in 5 sections ordered from $k = 1$ to $k = 5$ from left to right. The length of the kth section is proportional to $v_s[k]$, and the right-end of a section shows the cumulative proportion $\sum_{i=1}^{k} v_s[i]$ read on the linear-scale x-axis. The spatial ordering of the sections in a bar conveys the ordinal information about the k values, so we can use a qualitative color scheme for each section to better identify a specific degree k of lineage dissimilarity for cross-sample comparison. Sample names are given along the y-axis and with bars, are ordered according to the cumulative proportion of 1 to 5 within-sample lineage dissimilarities.

Rendering: The color code is a 5-class qualitative color scheme given by ColorBrewer [5].

Expected interpretation: In this view, discrepancy among the samples' lineage dissimilarity distributions should pop up through strong differences between lengths of same-color sections. More quantitative estimation of the discrepancy could be read on the x-axis scale.

3.4 Map of within-sample kinship similarities view

This view is designed to visualize both within-sample and between-sample kinship similarities at one glance.

Geometrical encoding: This view is a small-multiple representation of each sample as a scatterplot. The sample views are spatially organized based on some between-sample similarities. Each scatterplot represents the n_s individuals of a population sample s. We use MDS to visualize the kinship similarity matrices $\mathbf{M_s}$ with the objective that Euclidean distances in the projection space approximate $\sqrt{1-\mathbf{M_s}}$. We color the points in accordance with the color-coded bar graph view (3.3) to distinguish points i based on the lowest rounded lineage dissimilarity value to which they contribute: $d_{i*} = \min_x(d_{ix})$. Points with $d_{i*} > 5$ are plotted in black. We also connect pairs of points (i, j) with a straight line segment when their rounded lineage dissimilarity d_{ij} is in the set $\{1, 2, 3\}$ and color the line accordingly with the same color-code.

Rendering: Links are rendered before the points, and points or links are displayed by decreasing order of their lineage dissimilarity so more anomalous relationships are showed in front. Axes visibility, scale and aspect ratio follow the same rules as in view (3.2). Population codes (Table 2) are displayed in the low-density regions of the sample views to avoid clutter.

Expected interpretation: The between-sample anomalies are expected to pop up as links or distinct clusters appearing in some

Table 2: **Data characteristics**: Area codes: East Asia (EAS); Europe (EUR); Africa (AFR); South America (SAM); South Asia (SAS).

Area	Code	n_s	Description
EAS	CDX	93	Chinese Dai in Xishuangbanna, China
	CHB	103	Han Chinese in Bejing, China
	CHS	105	Southern Han Chinese
	JPT	104	Japanese in Tokyo, Japan
	KHV	99	Kinh in Ho Chi Minh City, Vietnam
EUR	CEU	99	Utah Res. with N & W Europ. Ancestry
	FIN	99	Finnish in Finland
	GBR	91	British in England and Scotland
	IBS	107	Iberian Population in Spain
	TSI	107	Toscani in Italia
AFR	ACB	96	African Caribbeans in Barbados
	ASW	61	Americ. of Afric. Ancestry in SW USA
	ESN	99	Esan in Nigeria
	GWD	113	Gambian in W Divisions in the Gambia
	LWK	99	Luhya in Webuye, Kenya
	MSL	85	Mende in Sierra Leone
	YRI	108	Yoruba in Ibadan, Nigeria
SAM	CLM	94	Colombians from Medellin, Colombia
	MXL	64	Mexic. Ancest. from Los Angeles USA
	PEL	85	Peruvians from Lima, Peru
	PUR	104	Puerto Ricans from Puerto Rico
SAS	BEB	86	Bengali from Bangladesh
	GIH	103	Gujarati Indian from Houston, Texas
	ITU	102	Indian Telugu from the UK
	PJL	96	Punjabi from Lahore, Pakistan
	STU	102	Sri Lankan Tamil from the UK

small views but not in others. The within-sample anomalies are expected to pop up through existence of links and color code of the points. The similarity between samples is not encoded in proportion to their relative proximity as in the view (3.2) but regarding the qualitative similarity of patterns appearing in their representative scatter plots.

4 Case Study: The 1000 Genomes Project Data

4.1 Data

The 1000 Genomes Project [12] data $S_{TGP} = \{s_1, \ldots, s_{26}\}$ aggregate 26 samples from various populations worldwide, each containing about 100 individuals. Table 2 shows their characteristics.

4.2 Kinship and lineage dissimilarity computation

The kinship matrix $\mathbf{M_s}$ is computed for each of the 26 population samples $s \in S_{TGP}$. Linkage disequilibrium (LD) regions in each population were removed before computing the kinship matrices. For removal of LD regions and computation of kinship matrices we used Plink software [8]. SNPs were pruned for LD regions by computing pairwise genotypic correlation with the window size of 50 SNPs, step size of 5 SNPs and R^2 threshold of 0.2. R^2 is the multiple correlation coefficient for a SNP being regressed on all other SNPs simultaneously [8]. The lineage dissimilarity matrix $\mathbf{D_s}$ is computed setting $C = 16$ because d_{ij} values (corresponding to non-zero m_{ij} values) lie in the range $[1, 15]$ across all samples.

4.3 Visual analysis

The **Between-sample lineage dissimilarities view** (3.2) of the 16-dimension v_s representation is shown in the figure 1. The leading two axes gathering more than 95.25% of the total variance, we consider that the cluster structure observed in this projection is faithful. We observe that samples CLM, PEL, MXL and PUR, as well as STU, ITU and PJL form two clusters distinct from the core cluster of samples. GIH, ASW, JPT, and to a lower extent CHB and CHS, appear as outliers.

This is a first hint that the population samples in the 1000 Genomes Project are not homogeneous in terms of within-sample kinship relations.

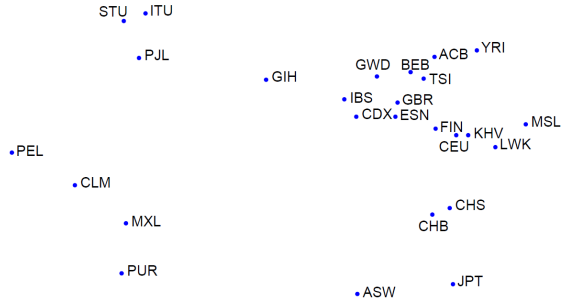

Figure 1: **Between-sample lineage dissimilarities view:** The MDS plot shows the population samples similarities based on the 16-dimension vector encoding of their lineage dissimilarity distributions. It shows a core cluster (right) of similar samples surrounded by samples of strongly different lineage characteristics. This overview shows an unexpected heterogeneity of lineage dissimilarity distribution across the samples, where {CLM, PEL, MXL, PUR} on bottom left, and {STU, ITU, PJL} on top left form two distinct clusters far apart the core one, and {ASW, JPT, GIH}, and {CHB, CHS} to a lower extent, appear as outliers.

Then we use the **distribution of lineage dissimilarities view** (3.3) to display (Figure 2) for each sample the distribution of its most anomalous lineage dissimilarities by increasing order of the proportions of values greater than 5. This visualization confirms that {PUR, PEL, CLM, MXL} form a cluster of samples having similar high proportion of lineage value 4, while additionally with {CHS, CHB, ASW, JPT}, they all have a higher cumulative proportion of anomalous lineage values 1 to 5. On the other hand {ITU, STU, PJL} have the lowest cumulative proportion of anomalous lineage values 1 to 5, which explains why they appear as clustered on the top left of the figure 1. This view gives more details about the anomaly distribution across the samples which confirm the global overview given in the figure 1.

At last we display the **map of within-sample kinship similarities view** (3.4) of each population sample in the figure 3. These MDS plots show that not only the lineage proportions are heterogeneous across samples as quantified by the bar graphs (Figure 2), but also that for the anomalous samples, clusters appear within the MDS plots while for the core set of homogeneous samples, no cluster or outlier appear confirming their homogeneous kinship content mostly made of lineage dissimilarities greater than 4.

Some of the detected samples that show higher relatedness of individuals within the population appear to either represent minorities in their residential areas, as in the case of MXL or ASW, or to belong to the South America area as for CLM, MXL, PEL and PUR. But we don't know yet about the causal relation behind this apparent correlation. These preliminary results call for a deeper analysis of the 1000 Genomes project data quality.

4.4 User feedback and guidelines

Users: The users are 3 experts in the domain of biology and genomic data analysis. A is a senior expert, while B and C are postdoctoral researchers. They are used to machine learning and standard data analysis techniques but not to exotic visualizations. Their endeavor was to discover new information from the kinship encoding of the 1000 Genomes Project data. They had no specific expectation of what would be discovered except that based on the general objective of the 1000 Genome Project, they all believed that the data should normally be homogeneous within and between the samples.

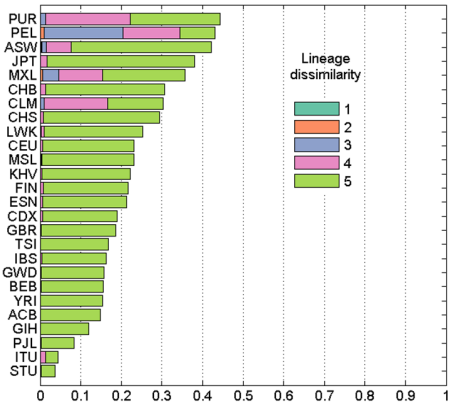

Figure 2: **Distribution of lineage dissimilarities view:** the proportion of anomalous lineage dissimilarities $v_s[k]$ for $k \in \{1,2,3,4,5\}$ are displayed for each sample as horizontal stacked bar graphs color-coded according to k. Bars are ordered from left to right for $k = 1 \ldots 5$ respectively. The samples are ordered on the y-axis based on their cumulative proportion from 1 to 5. PEL, CLM, PUR and MXL, clearly appear as having a significantly greater proportion of anomalous relations of value 4 (pink) than the other samples. PEL has a signigficantly greater proportion of value 3 (mauve). The top ranked PUR, PEL, ASW, JPT, MXL, CHB, CLM and CHS and bottom ranked STU, ITU, PJL and GIH samples also appear as clusters or outliers in figure 1.

Method: We showed the 3 visualizations to each user explaining the representations and answering their questions. Then we discussed with each of them in private to collect their feedback.

Feedback:

Between-sample view (Figure 1): B preferred this view showing heterogeneities at one glance through the existence of distinct clusters. C was not used to read MDS plots seeking axes meaning instead of focusing on points' proximities, and was not used to see each sample (set of individuals) represented as a single point.

Distribution view (Figure 2): A preferred this view clearly showing both heterogeneity of the samples and giving details about the distribution of their anomalous lineage dissimilarities. B found it also very useful to get a more detailed quantitative view. C was not sure about the meaning of the proportions displayed.

Map of within-sample view (Figure 3): C preferred this view showing strong relationships between individuals in each sample, as visible links whose absence in some plots clearly showed heterogeneities across samples. B raised questions regarding the clutter caused by many links possibly hiding information underneath, although the display ordering had been designed to limit this effect showing the most anomalous information on top. This view raised many questions among the users regarding the possible causes of heterogeneities calling for a deeper analysis of the data and the kinship processing.

MDS interpretation pitfalls: We verified that none of the users was aware of the mapping distortions [1] that could occur in MDS plots (Figures 1 and 3) . We checked that the MDS plots displayed were trustworthy due to the variance higher than 95% explained by their two principal axes, but we did not mention this fact to the users on purpose to *analyze how they can interpret these plots without direct knowledge of their trustworthiness*. When pointing them to the possible pitfalls known as false-neighbors [7], each of them told us having used a different strategy to convince herself about the trustworthiness of the view: A first observed that some of the population samples were located either close or far apart as she expected, which made her confident about the global quality of the plot, and then she analyzed the other pairwise distances as faithful representation of the original similarities and correctly detected the apparent heterogeneity as trustworthy; B said "I trust you, if you show me

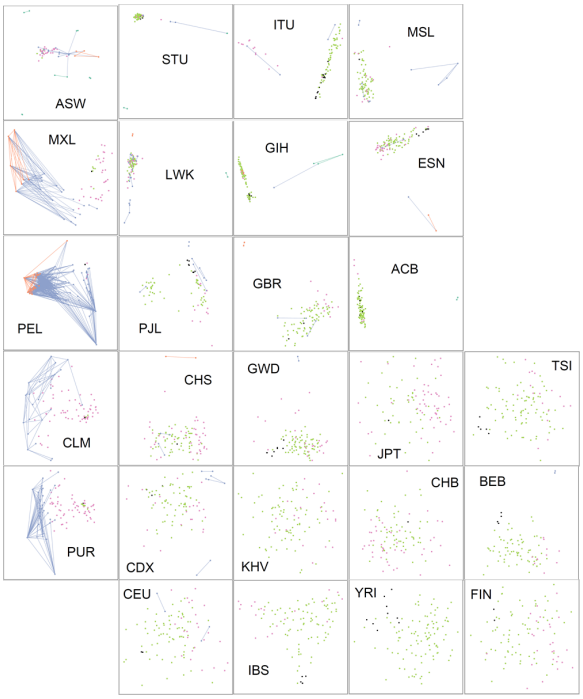

Figure 3: **Map of within-sample kinship similarities view:** each sample kinship similarity matrix is displayed as a color-coded scatterplot using MDS. The color-code is identical to the one used in figure 2 except that individuals with lineage dissimilarity greater than 5 are displayed as black dots. Color-coded links connect the dots involved in lineage dissimilarity up to 3. The scatterplots have been manually, tentatively and spatially organized by visual similarity according to their number of links (more links to the left), and to their number of clusters (more clusters to the top). The most homogeneous samples are in the lower right corner while the others call for a deeper quantitative analysis of their anomalous characteristics.

these figures, they must be correct"; C decided not to attempt to interprete the between-sample MDS plot she did not understand, still she could read the map of within-sample MDS plots focusing her attention on the qualitative differences between the subplots.

These latter outcomes regarding MDS plots interpretation show that MDS plots alone with no guidance could lead to false discoveries, but used together with additional trustworthy plots (here the bar graph) they might help to detect the unexpected (detecting heterogeneity while homogeneity is expected) and to focus the attention on the most interesting samples (here the clusters appearing in the MDS plot looked after in the bar graph plot).

5 Conclusion

We designed visualizations to enable the detection of anomalous kinship relations in Genome Wide Association Studies. Applying these visualizations in a case study based on kinship data from the 1000 Genomes Project, three expert users detected unexpected heterogeneities regarding the kinship relations. Some hypotheses have been drawn regarding these anomalies but beyond confirming the usefulness of these visualizations, they primarily call for a deeper analysis of the quality of the samples used in the 1000 Genomes Project as it may impact many projects which rely on these data.

Informal user feedback also provided us with possible ways to improve the visualizations: one is to give guidelines to interpret MDS plots and to better guard against their pitfalls, another is to automatize the spatial organization of the small multiples in the map of within-sample kinship similarities view using specific visual quality metrics [3,4,11,17].

REFERENCES

[1] M. Aupetit. Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing*, 70(7-9):1304–1330, 2007.

[2] M. Aupetit. Sanity check for class-coloring-based evaluation of dimension reduction techniques. pages 134–141, 2014.

[3] M. Aupetit and M. Sedlmair. Sepme: 2002 new visual separation measures. *Proceedings of IEEE PacificVis 2016 (this issue)*, 2016.

[4] E. Bertini and G. Santucci. Visual Quality Metrics. In *Proc. AVI Wkshp. BEyond time and errors: novel evaluation methods for information visualization (BELIV)*. ACM, 2006.

[5] C. Brewer, M. Harrower, B. Sheesley, A. Woodruff, and D. Heyman. Color brewer 2.0. *http://colorbrewer2.org*.

[6] J. Kruskal and M. Wish. *Multidimensional Scaling*. Sage, 1978.

[7] S. Lespinats and M. Aupetit. Checkviz: Sanity check and topological clues for linear and non-linear mappings. *Comput. Graph. Forum*, 30(1):113–125, 2011.

[8] A. L. Leutenegger, B. Prum, E. Génin, C. Verny, A. Lemainque, F. Clerget-Darpoux, and E. A. Thompson. Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet*, 73(3):516–523, 2003.

[9] M. D. Meyer, T. Munzner, and H. Pfister. Mizbee: A multiscale synteny browser. *IEEE Trans. Vis. Comput. Graph.*, 15(6):897–904, 2009.

[10] P. Oliehoek, J. J. Windig, J. A. M. van Arendonk, and P. Bijma. Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. *Genetics*, 173:483496, 2006.

[11] M. Sedlmair and M. Aupetit. Data-driven evaluation of visual quality measures. *Comput. Graph. Forum*, 34(3):201–210, 2015.

[12] Team. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[13] P. M. Van Raden. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423, 2008.

[14] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J.-D. Fekete, and D. W. Fellner. Visual analysis of large graphs: State-of-the-art and future research challenges. *Comput. Graph. Forum*, 30(6):1719–1749, 2011.

[15] R. E. Voorrips, M. C. A. M. Bink, and W. E. van de Weg. Pedimap: Software for the visualization of genetic and phenotypic data in pedigrees. *Journal of Heredity*, 103(6):903907, 2012.

[16] C. Ware. *Information Visualization - Perception for Design*. Morgan Kaufmann, San Francisco, 2004.

[17] L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: interactive exploration guided by pairwise views of point distributions. *IEEE Trans. on Visualization and Computer Graphics*, 12(6):1363–72, 2006.