Effective principal components analysis of SNP data

Hugh G. Gauch, Jr.[1¶*], Sheng Qian[2¶], Hans-Peter Piepho[3], Linda Zhou[1], and Rui Chen[2]

[1] Soil and Crop Sciences, College of Agriculture and Life Sciences, Cornell University, Ithaca, New York, United States of America

[2] Biological Statistics and Computational Biology, College of Agriculture and Life Sciences, Cornell University, Ithaca, New York, United States of America

[3] University of Hohenheim, Institute of Crop Science, Biostatistics Unit, Stuttgart, Germany.

[¶] These authors are joint first authors.

* hgg1@cornell.edu (HG)

Abstract

PCA is frequently used to discover and display patterns in SNP data from humans, animals, plants, and microbes—and especially to elucidate population structure. Given the popularity of PCA, one might expect that PCA is understood well and applied effectively. However, a survey of 125 representative articles from the literature reveals five departures from best practices. First, PCA works best with SNP coding 1 for the rare allele and 0 for the common allele, but many articles use some other coding. Second, we determined that double-centered PCA is the best variant of PCA for SNP data, but not a single article specified this choice. Third, SNP coding and PCA variant affect the quality and interpretation of PCA results, yet contrary to contemporary standards in science for transparency and reproducibility, few articles report these choices. Fourth, PCA biplots of both SNPs and Items (such as persons or cultivars) are more informative than PCA graphs of only Items, but not a single article presented a biplot. Fifth, PCA graphs and biplots should use the same scale for all axes, but only 10% of the articles used correct scaling. The recommended best practices for PCA of SNP data are easily implemented and may yield new insights, especially about causality. The lessons learned here for PCA analysis of SNP data have implications for other statistical analyses and other kinds of genomics data.

Author summary

Principal components analysis has been applied extensively to single nucleotide polymorphism data throughout the biological sciences, including applications in medicine and agriculture. But this process involves many details and choices, so we have asked a neglected but important question: Could PCA of SNP data be significantly improved? That question prompted us to determine best practices and to characterize contemporary practices with a literature survey. We found substantial deficiencies in contemporary practices. Fortunately, switching from contemporary to best practices is a simple matter of informed choices, rather than a laborious burden of more difficult analyses or extensive computations. Extraction of full value and benefit from SNP data can accelerate important applications in human medicine and in plant and animal breeding.

Introduction

Human, animal, plant, and microbial research increasingly relies on high-throughput sequencing technology to detect associations between interesting phenotypes and single nucleotide polymorphisms (SNPs). Principal components analysis (PCA) is commonly used to discover and display patterns in such data, especially to elucidate population structure. PubMed database queries return 59,000 hits for "principal component(s)", 135,000 hits for "SNP(s)", and 1,400 hits containing both terms (accessed 16 August 2018).

Given the popularity of PCA, it might be expected that PCA analysis is a well-understood technique that is routinely applied in the most effective way. However, evaluation of 125 examples from the literature reveals that this is not the case (S1 Table). In fact, researchers often appear to be unaware of PCA refinements that could enable them to extract considerably more information and value from their data. In this article, we identify shortcomings in prevalent PCA methodologies and describe several best practices that can provide researchers with stronger conclusions, especially about causal factors.

Best practices emerge here partly from comparing variants among PCA analyses of SNP data in the literature survey, partly from importing expertise from other contexts such as PCA analyses of ecological and agricultural data, partly from experiments with constructed and real datasets, partly from exploring statistical theory, and partly from inventing new tables and procedures presented here for the first time. Our recommendations are easily implemented for future experiments, and may yield new insights to those researchers who revisit datasets that have already been collected, analyzed, and published.

Results

Construction and interpretation of PCA graphs and biplots

The visualization of population structure is one of the most common applications of PCA to SNP data. PCA analyzes both matrix rows and columns [1]. The distinction between a PCA graph and a PCA biplot is that the former has points for only the rows or only the columns of a data matrix, whereas the latter includes both. In the present context, this means that a biplot has points for the Items *and* for the SNPs, where "Items" is a generic term for the samples, such as individual humans, horses, cultivars of wheat, or races of a pathogen. Obviously, a biplot is more informative than a graph, and only a biplot can reveal the important joint structure of the Items and SNPs. This is why biplots, after they were first introduced by Gabriel [1], have become the norm in countless applications of PCA [2]. Regrettably, our literature survey did not encounter even one biplot, so this powerful tool has been used in the context of PCA of SNP data only very rarely, if ever.

A potential hindrance to our advice to upgrade from PCA graphs to PCA biplots is that the SNPs are often so numerous that they would obscure the Items if both were graphed together. One way to reduce clutter, which is used in several figures in this article, is to present a biplot in two side-by-side panels, one for Items and one for SNPs. Another stratagem is to focus on a manageable subset of SNPs of particular interest and show only them in a biplot in order to avoid obscuring the Items. A later section on causal exploration by current methods mentions several procedures for identifying particularly relevant SNPs.

One of several data transformations is ordinarily applied to SNP data prior to PCA computations, such as centering by SNPs. These transformations make a huge difference in the appearance of PCA graphs or biplots. A SNPs-by-Items data matrix constitutes a two-way factorial design, so analysis of variance (ANOVA) recognizes three sources of variation: SNP main effects, Item main effects, and SNP-by-Item (S×I) interaction effects. Double-Centered PCA (DC-PCA) removes both main effects in order to focus on the remaining S×I interaction effects. The resulting PCs are called interaction principal components (IPCs), and are denoted by IPC1, IPC2, and so on. By way of preview, a later section on PCA variants argues that DC-PCA is best for SNP data. Surprisingly, our literature survey did not encounter even a single analysis identified as DC-PCA.

The axes in PCA graphs or biplots are often scaled to obtain a convenient shape, but actually the axes should have the same scale for many reasons emphasized recently by Malik and Piepho [3]. However, our literature survey found a correct ratio of 1 in only 10% of the articles, a slightly faulty ratio of the larger scale over the shorter scale within 1.1 in 12%, and a substantially faulty ratio above 2 in 16% with the worst cases being ratios of 31 and 44. Especially when the scale along one PCA axis is stretched by a factor of 2 or more relative to the other axis, the relationships among various points or clusters of points are distorted and easily misinterpreted. Also, 7% of the articles failed to show the scale on one or both PCA axes, which leaves readers with an impressionistic graph that cannot be reproduced without effort. The

contemporary literature on PCA of SNP data mostly violates the prohibition against stretching axes.

Each PC is associated with a singular value that equals the square root of the sum of squares (SS) captured by that PC, and this SS is called the eigenvalue. The singular value can be partitioned between the Item and SNP eigenvectors by the parameter α [3]. This article sets $\alpha = 0.5$. However, when the number of SNPs greatly exceeds the number of Items, a biplot using $\alpha = 0.5$ compresses the SNPs into a small region relative to the Items. Partitioning the singular value in favor of SNPs can produce a more satisfactory biplot, which means setting $1 > \alpha > 0.5$ for SNPs and $1-\alpha$ for Items.

The percentage of variation captured by each PC is often included in the axis labels of PCA graphs or biplots. In general this information is worth including, but there are two qualifications. First, these percentages need to be interpreted relative to the size of the data matrix because large datasets can capture a small percentage and yet still be effective. For example, for a large dataset with over 107,000 SNPs for over 6,000 persons, the first two components capture only 0.3693% and 0.117% of the variation, and yet the PCA graph shows clear structure (Fig 1A in [4]). Contrariwise, a PCA graph could capture a large percentage of the total variation, even 50% or more, but that would not guarantee that it will show evident structure in the data. Second, the interpretation of these percentages depends on exactly how the PCA analysis was conducted, as explained in a later section on PCA variants. Readers cannot meaningfully interpret the percentages of variation captured by PCA axes when authors fail to communicate which variant of PCA was used.

Two computational approaches to PCA are common. First, eigenanalysis or spectral decomposition is applied to a square and symmetric matrix, such as a SNPs-by-SNPs variance-covariance matrix that has been calculated from the SNPs-by-Items data matrix. Second, singular value decomposition (SVD) is applied directly to a SNPs-by-Items matrix, as explained in the appendix, ordinarily after a data transformation. SVD is used here because it provides a dual analysis of both SNPs and Items, and hence SVD is suitable for making biplots.

Enormous SNP datasets are becoming increasingly common, and fast PCA algorithms can readily handle large-scale genome-wide data. The remarkably efficient software FastPCA computes the top several PCs with time and memory costs that are linear in the number of matrix entries [5]. The software flashpca is also very fast [6]. The power method is the simplest algorithm for PCA and is efficient when only the first few PCs are needed [7]. This is the algorithm that we used here (S2 R Code). Another way to achieve computational tractability, that can be acceptable for some SNP datasets and research purposes, is to thin the SNPs prior to PCA, selecting the best SNPs by a tool such as PLINK [8].

The interpretive principles of PCA biplots merit a concise overview because biplots are virtually nonexistent in the genomics literature. This article focuses on DC-PCA biplots of SNP data, but readers desiring a more general account of biplots can consult the magisterial book by Gower et al. [2]. The underlying geometry is that Items (or SNPs) with similar data tend to be nearby in PCA graphs or biplots, whereas Items (or SNPs) with dissimilar data tend to be distant.

For two points of the same kind (both Items or else both SNPs), the cosine of the angle between them relative to the origin approximates their correlation: Two points in the same direction are positively correlated; two points in opposite directions are negatively correlated and hence indicate a contrast or gradient; and two points at right angles are uncorrelated. For two points of different kinds (an Item and a SNP), Items in a given direction have large positive S×I interactions for SNPs in that same direction, Items in a given direction have large negative S×I interactions for SNPs in the opposite direction, and Items in a given direction have small S×I interactions for SNPs at right angles. For points of the same or different kinds, the accuracy of these approximations increases with the percentage of the SS for S×I interactions that is captured in the IPC1-IPC2 plane. These interpretive principles apply more strongly to points near the periphery than to points near the origin. Two additional matters, concerning SNP coding and the so-called PCA arch distortion, are best deferred to the following two sections.

SNP coding

Consider a data matrix comprised of a number of SNPs observed for a number of Items such as persons or cultivars. The original reads of nucleotides (A, T, C, and G) must be coded numerically for PCA, such as a polymorphism of T and C being coded as 0 and 1. Three coding options for each SNP are compared in this section: code the rare allele as 1, code the common allele as 1, or assign codes in an arbitrary manner that does not distinguish between rare and common alleles. Apparently, the question has not yet been raised about how these choices of SNP coding affect PCA graphs and results.

The very popular variant call format (VCF) is an example of the third option. It assigns 0 to the allele of the reference genome and 1 to the non-reference genomes, and some additional details apply if there are more than 2 alleles [9]. Accordingly, the reference genome and others like it have an average over all SNPs of exactly or nearly 0, whereas extremely different genomes may have an average close to 1. Naturally, some SNPs in a reference genome are the rare allele and others the common allele, so there is a mixture in which both rare and common alleles are assigned 0, although common=0 dominates unless the reference genome is odd relative to everything else. The VCF format was developed for the 1000 Genomes Project in human genetics, but it has been adopted widely, so a large amount of SNP data has been stored in this format for humans, animals, plants, and microbes.

Fig 1 illustrates the first option, SNP coding rare=1. In the dataset shown on the left, the 24 columns represent the Items and the 20 rows represent the SNPs, with zeros denoted by dots in order to make the 1s readily visible. The convention adopted here is to number matrix columns from left to right and matrix rows from top to bottom, starting with 1 for the first column and the first row. This simple example is taken from a blogpost by Morrison et al. [10]; also see Morrison et al. [11]. The concentration of 1s along the matrix diagonal constitutes a single gradient with evident joint structure that involves *both* Items and SNPs.

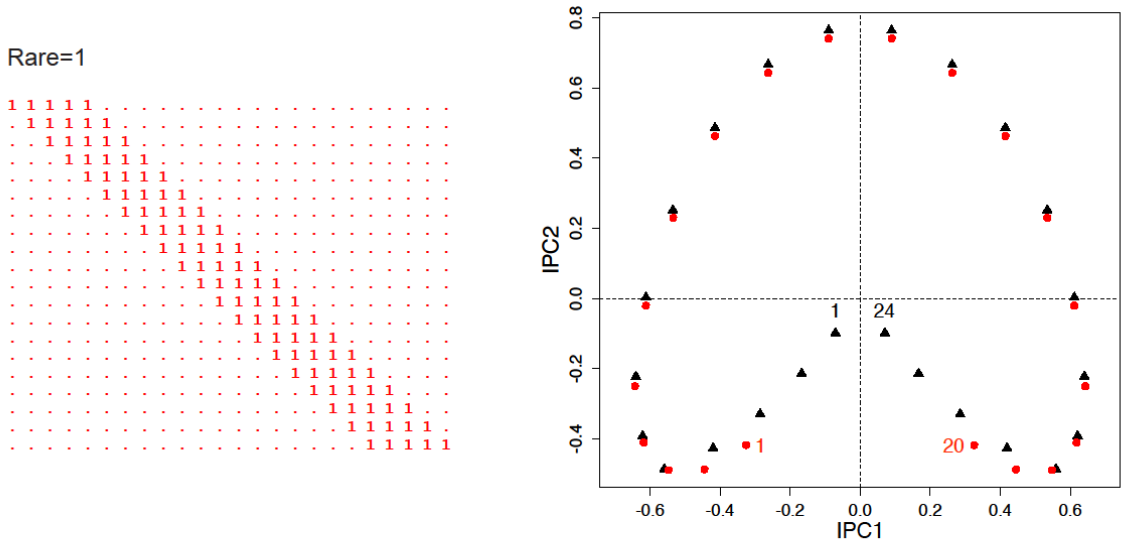

Fig 1. A simple gradient with SNP coding rare=1 and its DC-PCA biplot. The matrix has 24 Items in its columns and 20 SNPs in its rows. The biplot shows Items as black triangles and SNPs as red dots.

The DC-PCA biplot shown on the right in Fig 1 has two kinds of points: black triangles for the 24 Items, and red dots for the 20 SNPs. What should be a single gradient has been distorted into an arch with its ends involuted toward the middle for Items 1 to 24, and likewise for SNPs 1 to 20. Provided that one knows about this arch distortion, the single gradient is still apparent because these arches track each other in an intuitive manner that places related Items and SNPs in the same direction from the origin: Both arches move clockwise from Item 1 to 24 and from SNP 1 to 20. Ecologists have known for decades about this arch distortion, also called the horseshoe effect [12, 13]. This distortion also occurs with Nonmetric Multidimensional Scaling (MDS) and related methods [14]. The PCA arch has important implications for causal inferences, as will be explained in a later section on causal exploration by a new method.

However, an extensive search of the genomics literature by Morrison found only two papers that discuss the PCA arch distortion [15, 16]. Those papers and his blogposts have not yet succeeded in making this distortion well known (personal correspondence, David Morrison, 18 February 2018). Our survey found no additional mention of the arch distortion beyond the two papers that Morrison identified, even though about 80% of the PCA graphs had an evident arch, where by "evident" we mean obvious even at a mere glance. Researchers could interpret their PCA graphs more accurately were they aware of the arch distortion.

Only slightly more complex datasets than that in Fig 1 result in biplots that require some expertise for their proper interpretation. Fig 2 shows the same simple gradient as Fig 1, but with the second and opposite option, SNP coding common=1. Granted, the datasets in Figs 1 and 2

have exactly the same SS for everything: total, SNP main effects, Item main effects, S×I interaction effects, and every PC. Furthermore, the correlations are 1 or -1 between the opposite SNP coding of Figs 1 and 2 for both row and column PC scores, for all components, and for all six PCA variants (listed at the start of the next section). Those facts might prompt the expectation that reversing 0's and 1's makes no difference at all. But the DC-PCA biplot on the right in Fig 2 shows that such an expectation is false. The arch for Items 1 to 24 shown by black triangles and the arch for SNPs 1 to 20 shown by green dots both go clockwise. However, one arch is upside down relative to the other. Therefore, Item 1 and SNP 1 are opposite each other relative to the origin in Fig 2, although they were near each other in Fig 1, and the same applies to Item 24 and SNP 20. The reverse also obtains: Item 17 and SNP 6 nearly touch in Fig 2 (near the horizontal dashed line, at the right), although they were far apart in Fig 1. Consequently, Fig 2, unlike Fig 1, is counterintuitive and confusing because related Items and SNPS can be widely separated, and distant Items and SNPs can be quite close.

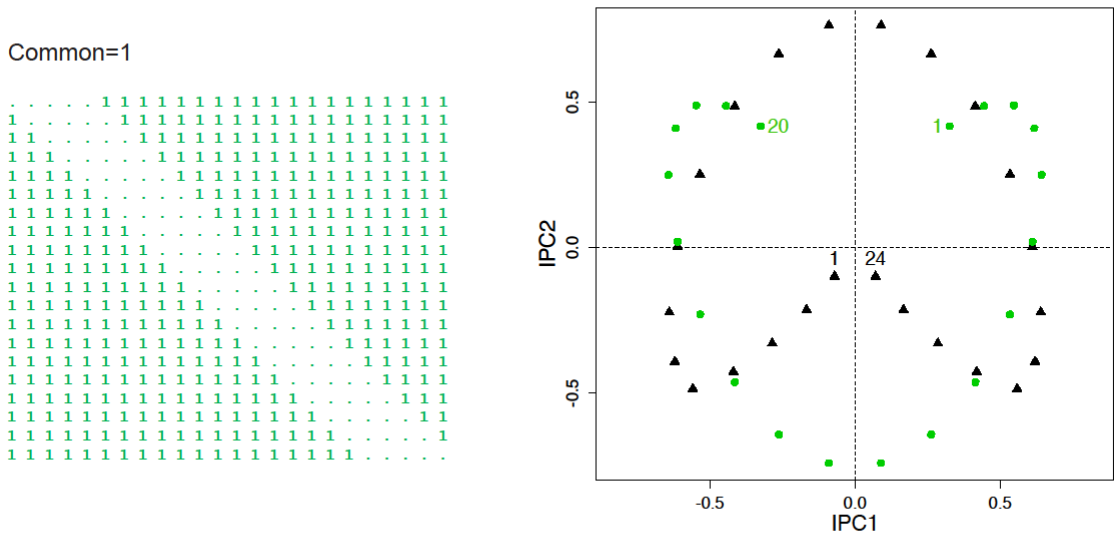

Fig 2. A simple gradient with SNP coding common=1 and its DC-PCA biplot. This is the same matrix as Fig 1 except for opposite SNP polarity. The biplot shows 24 Items as black triangles and 20 SNPs as green dots.

Fig 3 shows what happens with the third and final option, arbitrary SNP coding. In this dataset, the coding rare=1 and common=1 alternates, with rows using rare=1 shown in red, and those using common=1 shown in green. Incidentally, other arbitrary coding schemes give qualitatively the same results, such as selecting the coding at random for each SNP, or reversing the coding relative to that in Fig 1 for every fourth SNP—and the same would be expected for VCF. The DC-PCA biplot on the right combines the features shown previously in Figs 1 and 2. Items 1 to 24 are shown with black triangles, odd-numbered SNPs 1 to 19 with red dots, and

even-numbered SNPs 2 to 20 with green dots. All three arches go clockwise, but the orientation of the green arch is upside-down relative to the black and red arches. Taken together without distinguishing red from green, the dots for the SNPs roughly approximate a circle around the origin, rather than the typical arch, so awareness of the arch distortion would not be enough to guide proper interpretation. This biplot is quite confusing because SNPs near each other along the gradient can be far apart in the biplot (such as SNPs 1 and 2), whereas SNPs far apart from each other along the gradient can be near each other in the biplot (such as SNPs 2 and 13, where 13 is the red dot closest to green dot 2). Furthermore, this biplot inherits the problems already illustrated by Fig 2. Consequently, this biplot is doubly confusing.

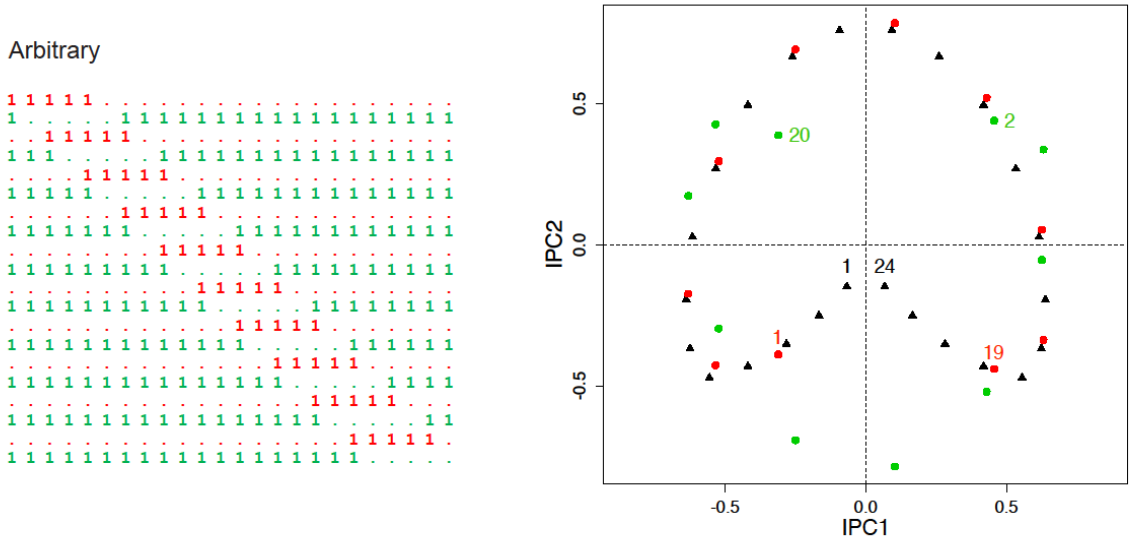

Fig 3. A simple gradient with arbitrary SNP coding and its DC-PCA biplot. This is the same matrix as Figs 1 and 2 except for arbitrary polarity, alternating SNP coding rare=1 (red) and common=1 (green). The biplot shows 24 Items as black triangles, 10 SNPs with coding rare=1 as red dots, and the remaining 10 SNPs with coding common=1 as green dots.

Although the DC-PCA biplots in Figs 1 to 3 show markedly different results for the SNPs (red or green dots), the results are exactly or nearly identical for the Items (black triangles). Indeed, correlations between IPC1 and IPC2 for Items are both exactly 1 for the datasets in Figs 1 and 2, and are 0.9999781 and 0.9983774 for the datasets in Figs 1 and 3. Fortunately, PCA graphs for Items are virtually immune to differences in SNP coding. The appendix explains this immunity and expands this investigation into SNP coding to all six PCA variants.

Fig 4 progresses to a new topic, the effect of different widths. The matrix on the left begins with the 20 SNPs of Fig 1, again shown in red, which have a uniform width of 5 consecutive 1s. Then 5 rows are added with different widths of consecutive 1s from 4 to 20, shown in blue. The DC-PCA biplot on the right is similar to Fig 1 for the 24 Items (black

triangles) and the first 20 SNPs (red dots), except for rotation of both by 90°. The numbers beside the blue points indicate widths (rather than SNP numbers as in the previous figures). As width increases from 4 to 20 for the blue SNPs in the last 5 rows of the matrix, the SNPs first migrate outward toward the periphery of the arch, and then they move inward toward the origin. Hence, highly informative SNPs (with many 1s and many 0s) are located around the periphery of the arch, whereas less informative SNPs (with rather few 1s or else very many 1s) are located near the middle of the arch. This is an important observation to bear in mind when interpreting PCA graphs or biplots that exhibit the arch distortion, and it actually imparts some interpretive value to that arch.

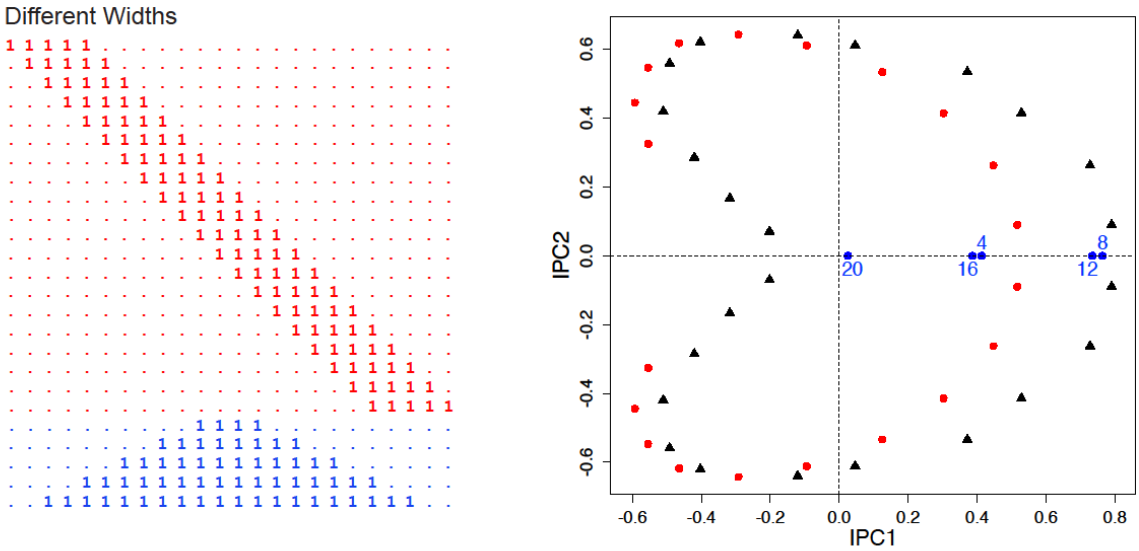

Fig 4. A simple gradient plus SNPs with different widths of distribution. The matrix has 24 Items in its columns and its first 20 rows are the same as Fig 1, with all of these rows having a width of 5. The additional 5 rows at the bottom, shown in blue, have widths from 4 to 20. The biplot shows the 24 Items as black triangles, the first 20 SNPs as red dots, and the additional rows as blue dots identified by their widths, 4, 8, 12, 16, and 20.

Fig 5 shows two constructed datasets with greater complexity that mimic a common situation: Items organized in several groups, such as African, European, and Asian persons. For both datasets, the columns represent 420 SNPs and the rows 25 Items, with values of 0 shown in white and values of 1 in black. The top dataset began with solid black blocks numbered 1 to 5 from left to right along the diagonal of the matrix, with some overlap, such as block 1 (SNPs 1–100) and block 2 (SNPs 80–180) that overlap in SNPs 80–100. Then noise was added by reversing 0/1 with probability 0.1. The top dataset uses the recommended SNP coding rare=1. The bottom dataset was generated from the top dataset by reversing the polarity for even-numbered SNPs, so it has arbitrary SNP polarity.

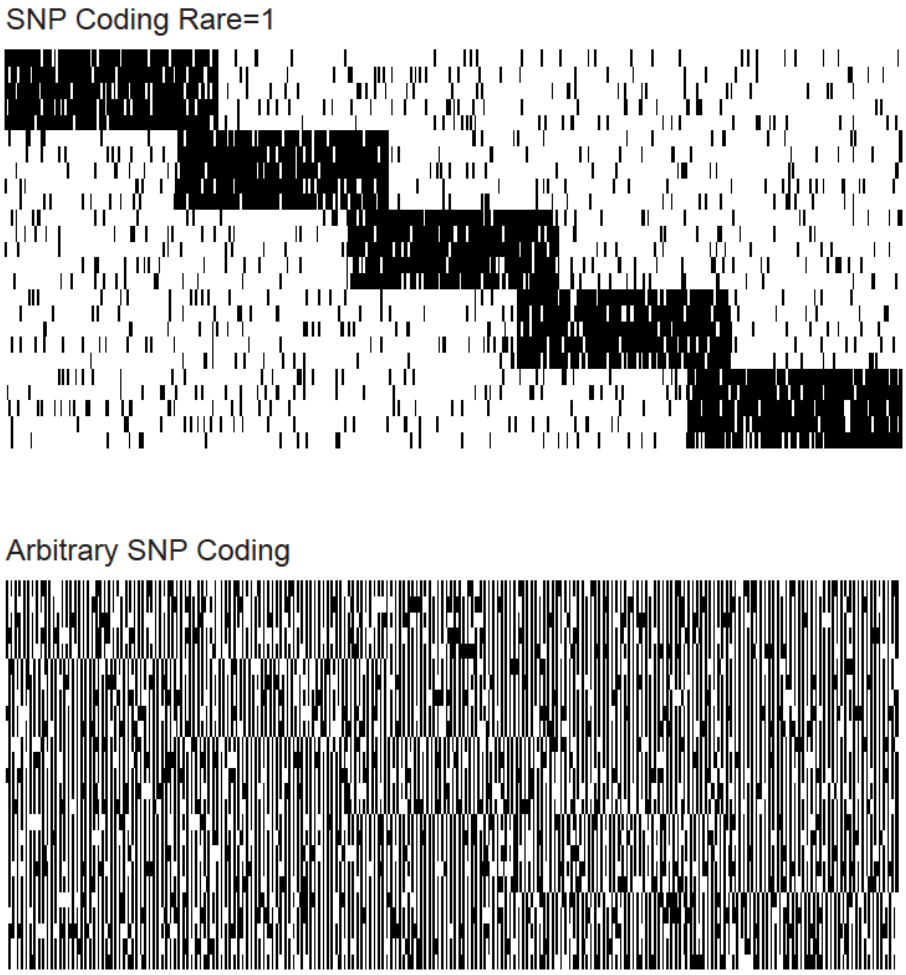

Fig 5. Constructed SNP datasets with SNP coding rare=1 (top) and arbitrary SNP coding (bottom).

Fig 6 shows the DC-PCA biplot for the dataset in the top of Fig 5 with SNP coding rare=1. It uses two panels for the biplot in order to avoid clutter. The Items show the typical arch distortion progressing in order from block 1 to 5 (red to green), with the 5 Items in each block tightly clustered. The SNPs show the same arch going from block 1 to 5. The SNPs that overlap between adjacent blocks are located somewhat further out toward the periphery, as expected for more informative SNPs. For example, the overlap shown with olive crosses is outside and in between blocks 1 and 2 shown with red and blue dots. The patterns for both Items and SNPs in this biplot are easy to interpret, provided that researchers are familiar with the arch distortion of PCA.

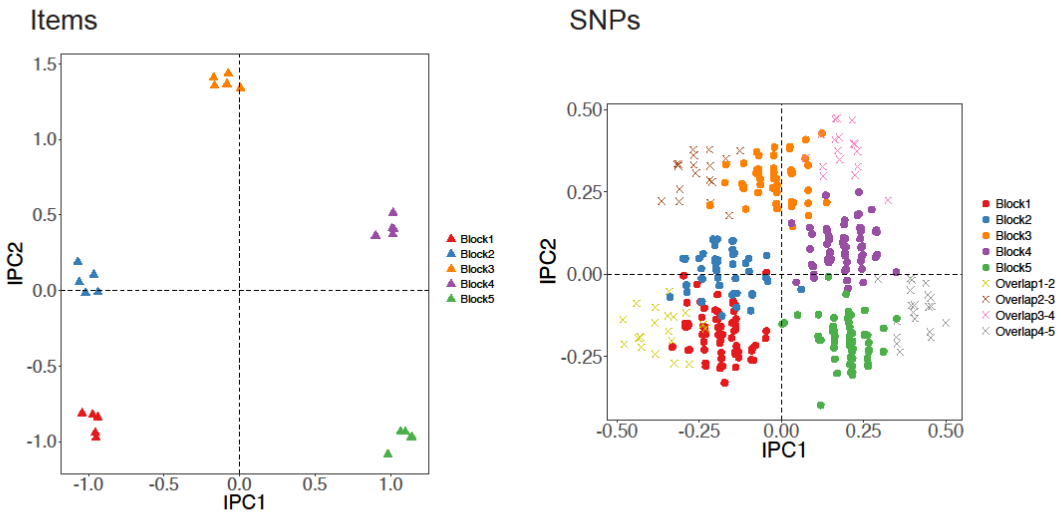

Fig 6. DC-PCA biplot for the constructed dataset with SNP coding rare=1.

Fig 7 shows the DC-PCA biplot for the arbitrary SNP polarity (bottom of Fig 5). Again, the Items on the left show the typical arch distortion. The Items in Figs 6 and 7 are visually identical because DC-PCA has virtual immunity to changes in SNP polarity, with the correlations being 0.9999996 for IPC1 and 0.9999986 for IPC2, and the appendix explains why a perfect correlation is not expected. However, each SNP block and overlap splits into two clusters located opposite each other relative to the origin (as could be anticipated from consideration of Fig 3). A PCA graph of actual data with these complicated features is quite difficult to interpret, even more so for researchers who are unaware of the PCA arch distortion.

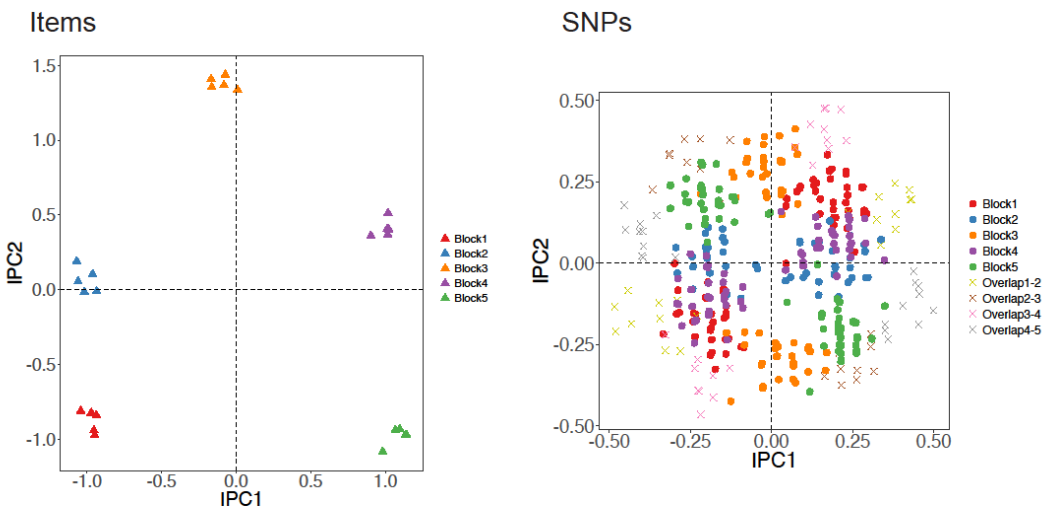

Fig 7. DC-PCA biplot for the constructed dataset with arbitrary SNP coding.

In review, the three choices for SNP coding, namely rare=1 or common=1 or arbitrary, make little difference for the Items portion of a DC-PCA biplot, whereas rare=1 is best for the SNPs portion because this choice enhances interpretability. Nevertheless, research purposes other than producing PCA graphs and biplots may call for different preferences for SNP coding. For example, when researchers are interested in the difference between reference and non-reference genomes, VCF coding makes the SNP main effects informative, whereas those effects are deliberately ignored by DC-PCA.

In our literature survey, specification of SNP coding was so rare that nothing substantial can be said about contemporary practices. On the other hand, specification of the software used to process SNP data is common, and that information may support an inference about the SNP coding, at least if the assumption is made that default options were selected. For example, the default option for TASSEL is SNP coding common=1 ([17]; Peter Bradbury personal communication, 23 April 2018). Nevertheless, inferences about coding based on software are far less than ideal, especially because numerous software packages are used, determination of the default SNP coding can be laborious, and selection of default options cannot be assumed for certain. Best practices require that articles include an explicit statement about which SNP coding was used, preferably accompanied by a reason for that choice that reflects an understanding of the consequences for subsequent PCA or other statistical analyses.

PCA variants

This section compares three PCA variants: SNP-Centered, Item-Centered, and Double-Centered PCA. They result from data transformations applied before PCA computations, namely removing main effects for SNPs or Items or both. Only DC-PCA was discussed in the previous

section because the considerations and results presented in the remainder of this article make it our default recommendation. This section also briefly mentions three additional variants: SNP-Standardized, Item-Standardized, and Grand-Mean-Centered PCA.

The primary consideration when choosing a PCA variant for analysis of SNP data is to identify those sources of variation that are relevant to a researcher's medical, agricultural, or other objectives. Again, a SNPs-by-Items dataset has three sources of variation: SNP main effects, Item main effects, and S×I interaction effects. The SNP main effects involve the average over Items for each SNP, and likewise the Item main effects involve the average over SNPs for each Item. Ordinarily these averages are not relevant to research objectives. Consequently, the differences between SNPs in different Items—that is, the S×I interactions—merit PCA analysis. This consideration alone justifies a default recommendation of DC-PCA for the analysis of SNP data.

We begin our comparison of PCA variants with some datasets already used in the previous section, the two constructed datasets in Fig 5. Table 1 shows ANOVA tables for DC-PCA of these two datasets with SNP coding rare=1 and arbitrary SNP coding. Sources are indented to indicate subtotals. Because DC-PCA removes both main effects prior to PCA calculations, PCA is applied to only the S×I interactions. The main difference caused by different SNP codings is that arbitrary coding increases the SS for the SNP main effects substantially, from 72.34057 to 559.61562. For any variant of PCA that removes the SNP main effects, including DC-PCA, that difference in SNP main effects is inconsequential. However, for any other variant that retains the SNP main effects, including Item-Centered PCA, that difference can have consequences.

Table 1. ANOVA table for DC-PCA of constructed datasets using SNP coding rare=1 and arbitrary SNP coding.

Source	df	SS Rare=1	SS Arbitrary
Total	10499	2137.70057	2624.97562
SNPs	419	72.34057	559.61562
Items	24	1.32438	1.76133
SxI	10056	2064.03562	2063.59867
IPC1	442	430.62358	430.60250
IPC2	440	366.15307	366.03974
IPC3	438	290.97965	290.97359
IPC4	436	247.67494	247.66975
IPC5	434	51.60778	51.63438
IPC6	432	48.79570	48.69304
IPC7	430	46.14681	46.06361
Residual	7004	582.05409	581.92207

Table 2 shows an ANOVA table for Item-Centered PCA of the dataset in the bottom of Fig 5 with arbitrary SNP coding. Item-Centered PCA removes only *one* of the main effects, the Item main effects. Accordingly, PCA is applied to the SNP main effects *and* S×I interaction effects combined (denoted by S&S×I), which has a SS of 559.61562 + 2063.59867 = 2623.21429. This combination of two sources of variation, unlike the single source in the previous table, requires a new approach in order to understand what is happening, namely an augmented ANOVA table that is introduced here for the first time. It partitions the SS of each PC into the portions due to main and interaction effects, as shown in its additional two columns. The required calculations are simple: For each PC, multiply its SNP scores and Item scores, which are a row vector and a column vector, to obtain the matrix of expected values, and then subject that matrix to ANOVA. Researchers who are familiar with PCA are accustomed to the automatic monotonic decrease in the SSs for the PCs, but note that the SSs for the SNPs and S×I portions in the last two columns are not monotonic.

Table 2. Augmented ANOVA table for Item-Centered PCA of a constructed dataset with arbitrary SNP polarity. PCA is applied to SNP main effects and S×I interaction effects combined (S&S×I), and the portion of each is shown in the last two columns.

Source	df	SS	SNPs	SxI
Total	10499	2624.97562		
Items	24	1.76133		
S&SxI	10475	2623.21429	559.61562	2063.59867
PC1	443	567.99963	549.12355	18.87608
PC2	441	430.28722	0.90572	429.38150
PC3	439	361.40194	7.97670	353.42524
PC4	437	290.92669	0.05015	290.87654
PC5	435	245.81490	1.46544	244.34945
PC6	433	51.60927	0.00233	51.60694
PC7	431	48.67771	0.00131	48.67640
Residual	7416	626.49694	0.09042	626.40652

As shown in Table 2, PC1 with its SS of 567.99963 is composed of 549.12355 for SNP main effects, which dominate, but only 18.87608 for S×I interaction effects. By contrast, PC2 captures only 0.90572 of SNP main effects, but 429.38150 of S×I interaction effects. The SNP main effects with a SS of 559.61562 dominate PC1 because this SS exceeds the first eigenvalue of the interaction 430.60250, even though this SS for the SNP main effects is far smaller than the SS of the entire interaction 2063.59867. Every PC contains some mixture of main and interaction effects.

Fig 8 shows the biplot for Item-Centered PCA for the data in the bottom of Fig 5 with arbitrary SNP coding, using separate panels for Items and SNPs to reduce clutter. On the left, PC1 for Items is unipolar, meaning that all values are of the same sign, in this case positive. This happens because PC1 is dominated by the SNP main effects, as Table 2 shows. By contrast, as Table 2 also shows, PC2 in this figure captures mostly S×I interactions. The biplot in Fig 8 is perplexing because of its mixture of mostly main effects in PC1 but mostly interaction effects in PC2, especially for anyone who does not produce the augmented ANOVA table and thereby become aware of the different kinds of information in each axis. Unlike the useful DC-PCA biplot in Fig 7, the Item-Centered PCA biplot fails to show the structure of this dataset.

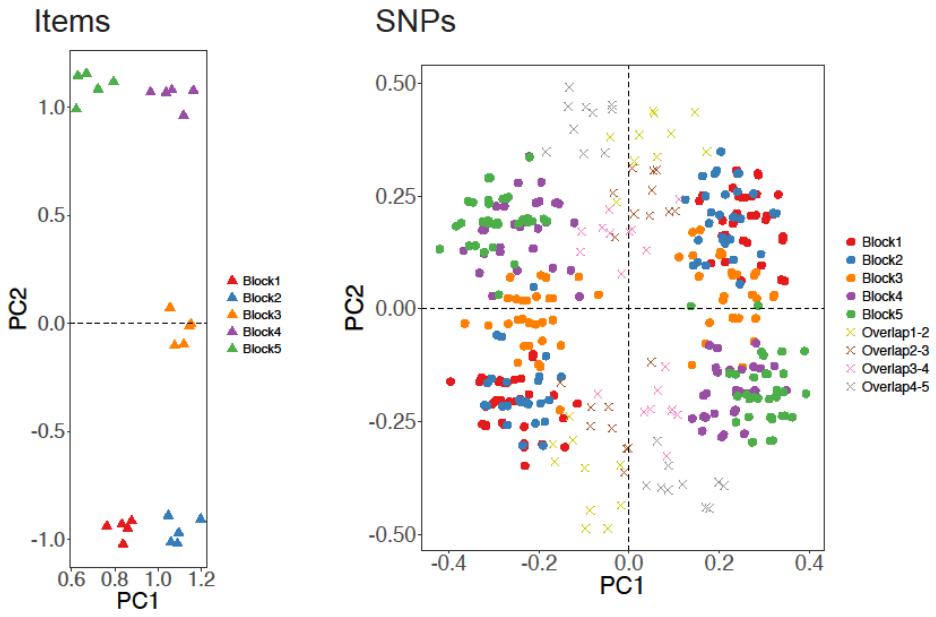

Fig 8. Item-Centered biplot for the constructed dataset with arbitrary SNP coding.

A PC that is unipolar for Items is diagnostic for a PC dominated by SNP main effects, and likewise a PC that is unipolar for SNPs is diagnostic for a PC dominated by Item main effects. When graphed properly with the same scale on every axis, PCA graphs with a unipolar axis tend to have a width-to-height ratio far from 1:1. For instance, the graph for Items on the left of Fig 8 is narrow, about 4 times higher than wide. Severe aspect ratios should be perceived as a warning to switch from an inappropriate variant of PCA to a more sensible variant.

Table 3 shows the augmented ANOVA table for Item-Centered PCA of the dataset in the top of Fig 5 with SNP coding rare=1. Comparing the results in Table 2 for arbitrary SNP coding and Table 3 for SNP coding rare=1, the decrease in SNP main effects from 559.61562 to 72.34057 demotes the SNP main effects from PC1 to PC5. Other schemes for SNP coding could cause other PCs to be dominated by the SNP main effects, such as PC2 or PC3. For any variant of PCA that retains the SNP main effects, including Item-Centered PCA, the choice of SNP coding can influence the SS for the SNP main effects, and thereby alter which PC is dominated by SNP main effects, which in turn changes how a PCA graph or biplot should be interpreted.

Table 3. Augmented ANOVA table for Item-Centered PCA of a constructed dataset with SNP coding rare=1. PCA is applied to SNP main effects and S×I interaction effects combined (S&S×I), and the portion of each is shown in the last two columns.

Source	df	SS	SNPs	SxI
Total	10499	2137.70057		
Items	24	1.32438		
S&SxI	10475	2136.37619	72.34057	2064.03562
PC1	443	430.73404	0.13879	430.59525
PC2	441	369.10712	3.66986	365.43726
PC3	439	291.04807	0.08956	290.95851
PC4	437	249.51934	2.46543	247.05392
PC5	435	70.70475	62.53384	8.17091
PC6	433	51.47682	0.26690	51.20992
PC7	431	48.72663	0.11631	48.61032
Residual	7416	625.05941	3.05987	621.99954

The principles shown above with constructed datasets can be reinforced and extended with real data on oats (*Avena sativa* L.) [18]. Kathy Esvelt Klos kindly shared with us a dataset with 635 lines of oats by 1341 SNPs that was used for PCA in their Fig 1 (personal correspondence, 4 June 2018). There are no missing data, every SNP is biallelic, and the two alleles were assigned values of 1 and 2, which we transposed to 0 and 1. The SNP data as received had arbitrary SNP coding. In order to have our recommended SNP coding rare=1, the polarity was reversed for 772 of the 1341 SNPs.

Table 4 shows the ANOVA table for DC-PCA of the oat data. The SS of 2077.80776 for Item main effects is smaller than 15566.72278 for IPC1, and that has implications for SNP-Centered PCA. By contrast, the SS of 16872.73111 for SNP main effects is larger than IPC1, and that has implications for Item-Centered PCA. Both of these variants are discussed next. The recommended DC-PCA removes both Item and SNP main effects and then applies PCA to a single source, S×I, so there is no need for an augmented table here.

Table 4. ANOVA table for DC-PCA of SNP data on oats.

Source	df	SS
Total	851534	157442.75631
SNPs	1340	16872.73111
Items	634	2077.80776
SxI	849560	138492.21743
IPC1	1973	15566.72278
IPC2	1971	9512.80403
IPC3	1969	5836.91490
IPC4	1967	4831.28663
IPC5	1965	3507.23659
IPC6	1963	2997.35142
IPC7	1961	2887.57428
Residual	835791	93352.32680

Table 5 shows the augmented ANOVA table for SNP-Centered PCA of the same oat data. This variant of PCA, which is not recommended, removes SNP main effects and then applies PCA to the Item main effects and S×I interaction effects combined (I&S×I). The Item main effects are so small that all 7 PCs and the residual are dominated by S×I interaction effects, as could be expected from Table 4. Indeed, the Item main effects account for only 4.3% of the SS captured in a PC1-PC2 graph. Were a biplot produced for this SNP-Centered PCA analysis, the intermixed 95.7% of S×I interaction effects would obliterate any possibility of perceiving structure or patterns due to Item main effects.

Table 5. Augmented ANOVA table for SNP-Centered PCA of SNP data on oats. PCA is applied to Item main effects and S×I interaction effects combined (I&S×I), and the portion of each is shown in the last two columns.

Source	df	SS	Items	SxI
Total	851534	157442.75631		
SNPs	1340	16872.73111		
I&SxI	850194	140570.02520	2077.80776	138492.21743
PC1	1974	16325.61203	857.67605	15467.93598
PC2	1972	9751.15835	257.21730	9493.94105
PC3	1970	6250.50356	423.13351	5827.37005
PC4	1968	4860.91024	23.38556	4837.52468
PC5	1966	3510.31175	4.52061	3505.79113
PC6	1964	3181.95302	187.61462	2994.33840
PC7	1962	2891.65618	1.65424	2890.00194
Residual	836418	93797.92007	322.60586	93475.31421

The argument against SNP-Centered PCA has three cases that exhaust the possibilities. First, if the Item main effects are not of interest, as is often the case, then they should also be removed, thereby resulting in the recommended DC-PCA. Second, if the Item main effects are of interest and the SS for these effects is small relative to the SS for PC1 and PC2—as happens for the oat example of Table 5 that is discussed in the previous paragraph—then a biplot using SNP-Centered PCA is wholly ineffective for displaying Item main effects. Third, if the Item main effects are of interest and their SS is large—as can happen with VCF or other biologically informative SNP codings—then Item main effects can dominate PC1 or PC2 and thereby produce a PCA graph or biplot with a confusing mixture of main and interaction effects. In no case is there any reason to choose SNP-Centered PCA.

The third case can be handled well by a modification of DC-PCA that includes main effects: an AMMI1 biplot, which is unknown in genomics but commonplace in the literature on agricultural yield trials. An AMMI1 biplot shows both of the main effects on the abscissa, and IPC1 on the ordinate. It captures 100% of both main effects in its abscissa, whereas SNP-Centered PCA necessarily captures less of the Item main effects and does not show SNP main effects at all. Also, IPC1 is the unique least-squares solution that captures as much of the S×I

interaction effects as possible, whereas SNP-Centered PCA necessarily captures less in its PC1. Furthermore, AMMI1 has no confounding: An AMMI1 biplot always captures only main effects in its abscissa, and only interaction effects in its ordinate. By contrast, SNP-Centered PCA mixes main and interaction effects in every PC, as an augmented ANOVA table demonstrates. Further information on the AMMI model and AMMI1 biplot is beyond the scope of this article, but is available in Gauch et al. [19].

Table 6 shows the augmented ANOVA table for Item-Centered PCA of the same oat data. This variant of PCA, which also is not recommended, removes Item main effects and then applies PCA to the SNP main effects and S×I interaction effects combined (S&S×I). This table shows that PC1 is dominated by SNP main effects (96.2%) whereas PC2 is dominated by S×I interaction effects (99.9%), as could be expected from Table 4. But that is only one possibility; the outcome for other datasets could be the reverse, or the SNP main effects might be small enough to be deferred to a higher component such as PC3 or PC4.

Table 6. Augmented ANOVA table for Item-Centered PCA of SNP data on oats. PCA is applied to SNP main effects and S×I interaction effects combined (S&S×I), and the portion of each is shown in the last two columns.

Source	df	SS	SNPs	SxI
Total	851534	157442.75631		
Items	634	2077.80776		
S&SxI	850900	155364.94855	16872.73111	138492.21743
PC1	1974	17402.41995	16734.51347	667.90648
PC2	1972	15564.01757	21.91820	15542.09937
PC3	1970	9476.48864	40.28319	9436.20545
PC4	1968	5781.80866	23.61243	5758.19623
PC5	1966	4758.74670	25.74541	4733.00129
PC6	1964	3482.96940	5.45197	3477.51743
PC7	1962	2970.19416	4.89492	2965.26988
Residual	837124	95928.30346	16.31151	95912.02131

The argument against Item-Centered PCA has exactly the same form as the argument against SNP-Centered PCA. In no case is there any reason to choose either of them.

Three additional variants of PCA were listed at the outset of this section: SNP-Standardized, Item-Standardized, and Grand-Mean-Centered PCA. All of them fail to resolve the underlying problem of mixed main and interaction effects, so they receive no further attention. For the oat example, the portion of main effects in the matrix submitted to PCA analysis is 1.5% for SNP-Centered PCA and 1.7% for SNP-Standardized PCA, with the remainder being interaction effects; and likewise the portion is 10.9% for Item-Centered PCA and 11.0% for Item-Standardized PCA. For Grand-Mean-Centered PCA, the portions are 1.3% for Item main effects, 10.7% for SNP main effects, and 88.0% for S×I interaction effects.

In review, researchers who use any variant of PCA other than DC-PCA should produce and publish an augmented ANOVA table in order to determine and communicate what sort of information is in each PC. Both SNP coding and PCA variant can affect which PC, if any, is dominated by main effects.

Our literature survey found specification of the PCA variant so rarely that nothing substantial can be said about contemporary practices. Unfortunately, the genomics community evinces little awareness that the choice of PCA variant has substantial implications for interpreting PCA graphs and for reproducing the PCA results of other researchers. Regrettably, our literature survey did not encounter even a single specification of DC-PCA, so applications of our recommended PCA variant for analysis of SNP data must be quite rare. Best practices require that articles include an explicit statement about which PCA variant was used, preferably accompanied by a reason for that choice.

Causal explanation by current methods

Discovering the causes of events is a fundamental goal of science [20] and SNP-based research is no exception [21, 22]. For example, migration, isolation, and admixture of populations cause changes in SNP frequencies that can be interpreted by PCA graphs [23, 24]. Demographic history can be reconstructed by focusing on SNPs that are correlated with PC1 and PC2 [25, 26]. These historical causal factors can be validated with independent data [27]. Natural selection and artificial selection (domestication) cause a genetic imprint on SNPs that can be elucidated by PCA graphs [28, 29]. Geographic separation causes genetic differentiation, so genes can mirror geography, as shown by figures that combine a PCA graph and a map [30-34].

Manhattan plots based on PC scores can be used to identify SNPs and thereby genes that are under strong selection within a given population. For example, Duforet-Frebourg et al. used several simulated datasets and two large human datasets to detect genomic signatures of natural selection with PCA [35]. They constructed Manhattan plots using PC1 and PC2 and found several well-known genes involved in local adaptation, as well as several new candidate regions meriting further investigation. Three of those authors produced the R package *pcadapt* that performs genome scans for genes under selection using PCA [36]. For another example,

Manhattan plots based on PC1 through PC4 identified 5 distinct subpopulations of European Americans, confirmed several previously known loci under selection, and found 3 novel loci [5].

Another way to identify SNPs that have causal significance is to focus on ancestry-informative markers (AIMs). For example, a panel of merely 23 AIMs, selected from datasets with over 1,000,000 SNPs, suffices to produce a PCA graph that has clear clusters for several major US populations [37, 38]. Of course, ancestry is just one particular instance of a causal factor. Informative SNPs could also be sought that reflect other sorts of causal factors or gradients, such as geography informative SNPs.

However, PCA offers additional possibilities for exploring causality.

Causal exploration by a new method

Recall that our survey of 125 articles on SNP data found that about 80% of the PCA graphs show an evident arch. That fact is highly relevant in the search for causal explanations. *Three things go together, necessarily and inseparably: a major causal factor or gradient that imposes joint structure on the rows and columns of a data matrix, a data matrix that can be arranged to concentrate large values along its diagonal, and a PCA arch*. This is one story told three ways; given any one of these things, all three will occur. Although this unified story has been familiar to ecologists for decades (Figs 3.1–3.2, Table 1.1, and Fig 4.7 in [12]), its relevance for SNP research has not yet been noticed.

Why not yet? Fig 9 shows how an obvious gradient involving the joint structure of SNPs and Items can become obscure, and it also shows how to recover the obvious joint structure. At the left is the simple gradient that was also shown in Fig 1, with 24 Items in the columns and 20 SNPs in the rows. The concentration of 1s along the matrix diagonal constitutes evident joint structure. However, as shown by the arrows pointing to the right in Fig 9, two problems can obscure the joint structure.

Fig 9. Origin of an obscure gradient and recovery of the original obvious gradient. Random ordering of the rows and columns of a matrix, combined with arbitrary SNP polarity, turns an obvious gradient into an obscure gradient. Contrariwise, SNP coding rare=1, followed by CA1 ordering of matrix rows and columns, recovers the original obvious gradient.

First, the Items and SNPs are commonly presented in a random order relative to the underlying causal/ecological/biological/historical gradient, as shown by the matrix in the middle. This matrix was constructed by shuffling the rows and then shuffling the columns of the matrix at the left. Both matrices have absolutely identical data, and they produce absolutely identical PCA graphs and ANOVA tables. Nevertheless, the order in which rows and columns are presented in those two matrices has a huge impact on our ability to readily grasp structure in the data.

Second, SNP polarity is often unrelated to the underlying causal gradient, as shown by the matrix on the right, which combines both problems: random order of rows and columns, and arbitrary SNP polarity. This matrix was constructed by reversing the polarity of the even-numbered rows of the matrix in the middle. The gradient or joint structure is as obvious in the matrix at the left as it is obscure in the matrix at the right.

Even as two problems turn an obvious gradient into an obscure gradient, two steps can recover the obvious joint structure, as shown by the arrows pointing to the left—at least for this special case of a single dominant gradient. First, use the recommended SNP coding rare=1. Second, order the matrix rows and columns by the rank order of the first component scores of correspondence analysis (CA1). CA is related to PCA and also involves SVD, but it uses chi-squared distances rather than Euclidean distances [13]. Biallelic data are sometimes coded as –1 and 1, but such data would need to be re-coded with non-negative values such as 0 and 1 because CA requires non-negative data. If as occasionally happens the arch is not oriented such that CA1 sweeps across the arch from one end to the other, then construct a linear combination of CA1 and CA2 that does sweep across the arch.

Fig 10 shows why the first step, SNP coding rare=1, is necessary. Recall that Figs 1 to 3 showed datasets and biplots for three choices of SNP coding: rare=1, common=1, and arbitrary. Fig 10 shows the results when the datasets from Figs 1 to 3 are arranged according to ranked CA1 scores for both the rows and the columns. CA attempts to concentrate 1s along the matrix diagonal (Fig 4.9 in [12]). But only the coding rare=1 allows ranked CA1 scores to recover the original order of the SNPs and Items, and thereby to display clearly the joint structure in this dataset. This is another reason, beyond others already given, for preferring the SNP coding rare=1.

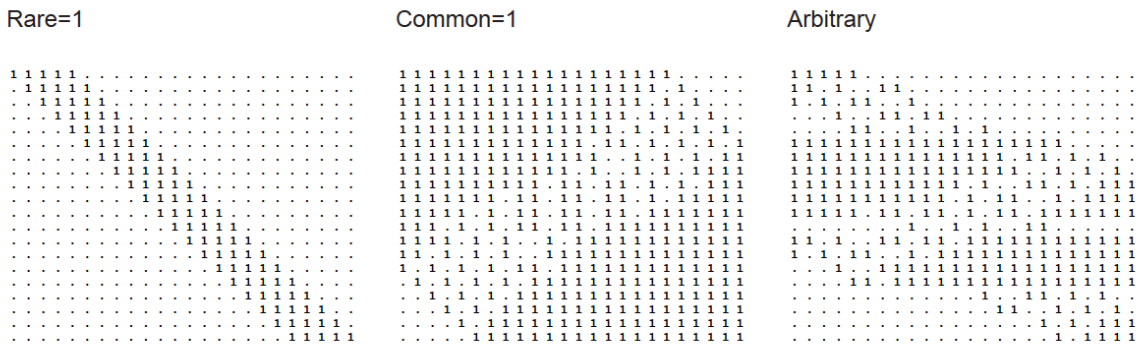

Fig 10. CA-arranged matrices for SNP codings rare=1, common=1, and arbitrary.

Fig 11 shows why the second step, order the matrix rows and columns by the rank order of CA1 scores, works only when given the first step, use SNP coding rare=1. Again we are using the datasets from Figs 1 to 3. In the left biplot with the recommended SNP coding, the Items and SNPs maintain their original order. However, in the middle biplot with SNP coding common=1 and the right biplot with arbitrary SNP coding, CA has involuted arches that alter the rank order of Items and SNPs along CA1.

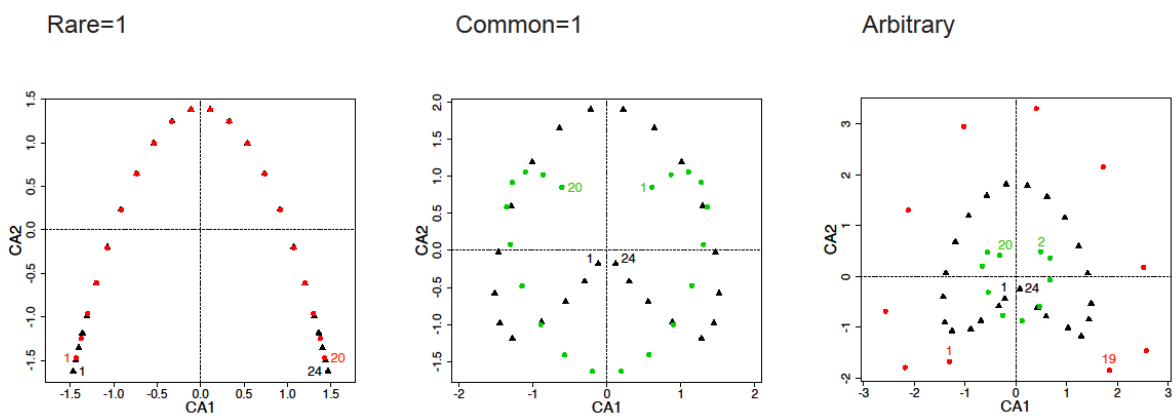

Fig 11. CA biplots for SNP codings rare=1, common=1, and arbitrary. The color schemes are the same as those in Figs 1 to 3.

Because this two-step procedure for recovering an obvious gradient works only for the special case of a single dominant gradient, the practical question arises: In the literature on PCA analysis of SNP data, how often are datasets structured substantially by a single dominant gradient? Given the one story told three ways, that question is equivalent to: How often do PCA graphs of the Items have an evident arch distortion? Again, the answer is about 80%. Many arches are a plain arch; but some are a filled arch, meaning that the arch is thick and includes

many points near the origin. The cause of a filled arch is wide variation in the number of 1s for each SNP, as was explored in Fig 4.

To illustrate gradient recovery using actual data, we deliberately selected an example of a heavily filled arch that is not as ideal as a plain arch, namely the oat dataset used above. Fig 12 shows the DC-PCA graph for these 635 oat lines, using SNP coding rare=1. From Table 4, IPC1 captures 11.2% of the S×I interactions and IPC2 captures 6.9% for a total of 18.1%. There are three groups of oats: 411 spring oats shown in green, 121 Southern US oats shown in red, and 103 world diversity oats shown in blue. This dataset is "weakly structured" and these three groups form "overlapping and diffuse clusters" [18]. The world diversity oats (blue) are most diffuse, as would be expected from their great diversity. The Southern US oats (red) are concentrated at the right, whereas the spring oats (green) are concentrated at the left. This graph is representative of the overwhelmingly most common application of PCA to SNP data: a graph of only Items, with the Items shown in several colors to display population structure that corresponds to a causal story involving geography, natural or artificial selection, medical conditions, or other causal factors.

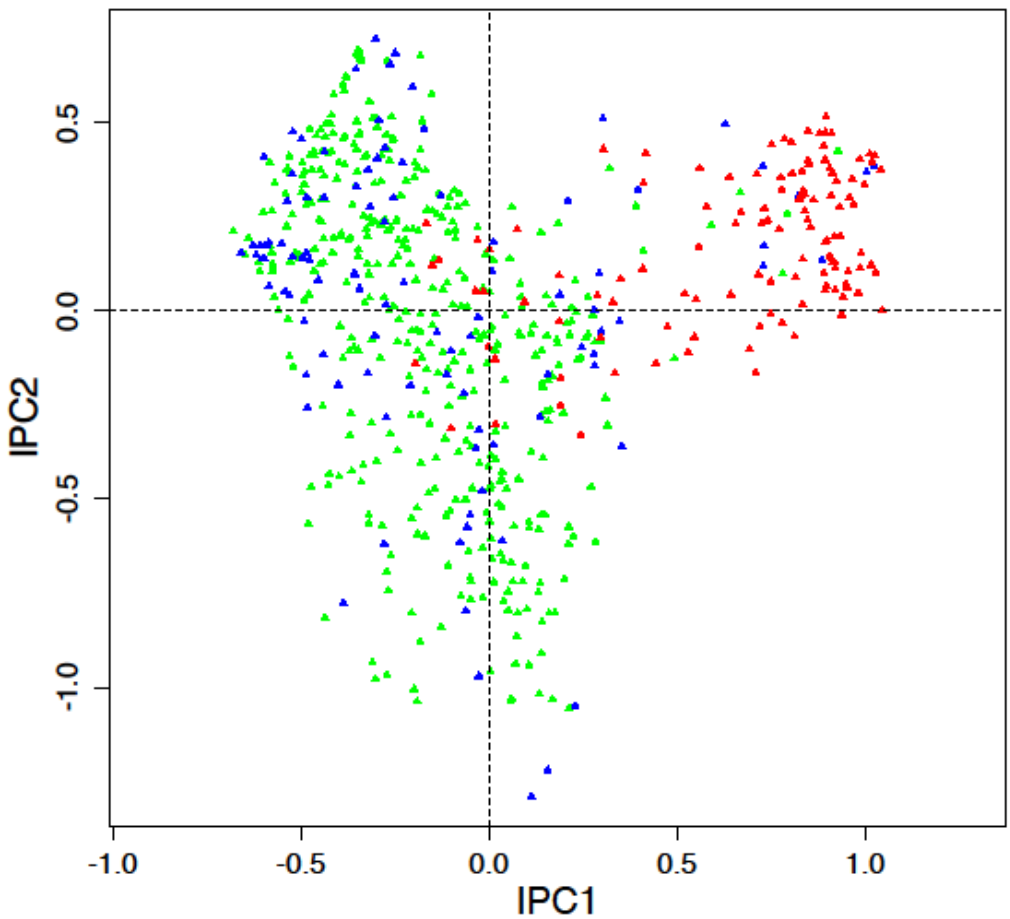

Fig 12. DC-PCA graph for 635 oat lines. The 411 spring oats are shown in green, the 121 Southern US oats in red, and the 103 world diversity panel oats in blue.

Recall from Fig 4 that a PCA arch has interpretive value: Points around the periphery of the arch are especially informative, whereas points near the middle are less informative. For example, the bottom two points in Fig 12, oat lines Ogle and Chaps, are specifically adapted for agroecological conditions in the middle of the spectrum from spring to Southern US oats. By contrast, the two points closest to the origin, lines Akiyutaka and HA05AB20-1, are rather unresponsive to the causal factors that distinguish spring and Southern US ecotypes.

Fig 12 has some interesting exceptions that invite further examination. For instance, line CI8000-4 was classified as a spring oat, but it appears as the right-most green point in the midst of numerous red points for Southern US oats. Likewise, line UPFA_22_Temprana was classified as a Southern US oat, but it appears as the left-most red point amidst mostly green points for spring oats.

Fig 13 shows the data with 635 oat lines in rows and 1341 SNPs in columns, using black=1 and white=0. The two steps, SNP coding rare=1 and CA-arranged matrix, make the structure evident, with 1s concentrated along the matrix diagonal. By contrast, the data as originally received—with arbitrary SNP polarity and random ordering of matrix rows and columns—shows no structure (data not shown). A simple proof that the structure in Fig 13 reflects a real causal gradient, rather than arises as an artifact of the CA1 ordering, is that randomizing the order of the oat lines for each SNP individually and then repeating the CA1 ordering makes the matrix structure disappear (data not shown).

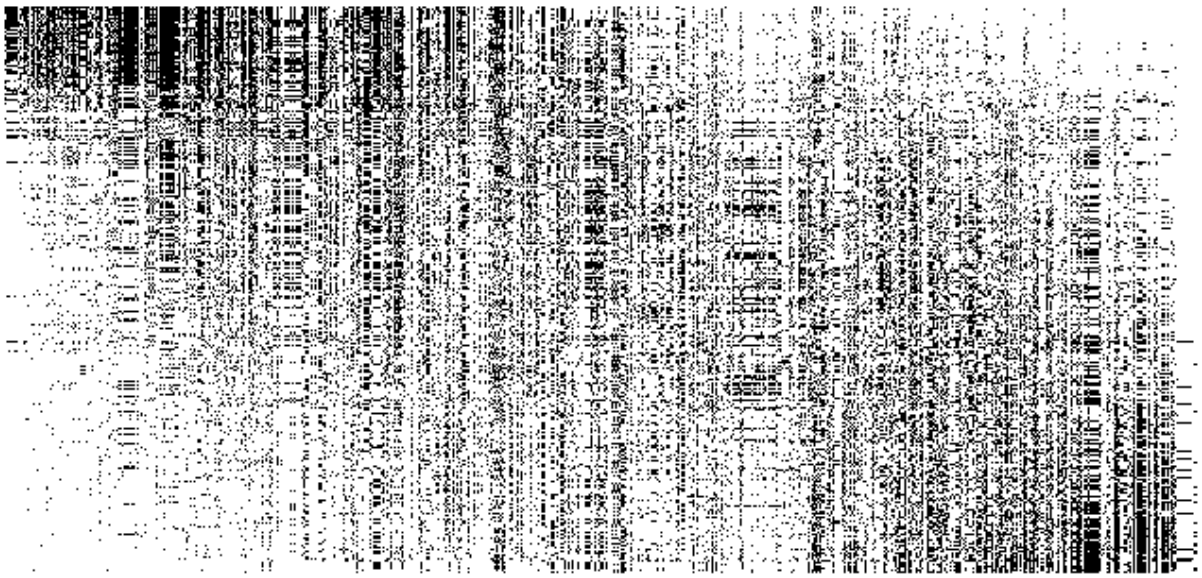

Fig 13. CA-arranged matrix for 635 oat lines in rows and 1341 SNPs in columns.

Fig 14 shows the DC-PCA biplot for the oat data, with colors derived from the CA order in Fig 13. The 635 lines are subdivided into 5 equal groups of 127 lines according to CA1 order from top to bottom in Fig 13, and likewise the 1341 SNPs form 5 groups of 268 SNPs (plus 1 extra for the last group) from left to right. The corresponding color scheme is: red, pink, black, light green, dark green. For example, red triangles in the left graph are oat lines 1–127, and red dots in the right graph are SNPs 1–268.

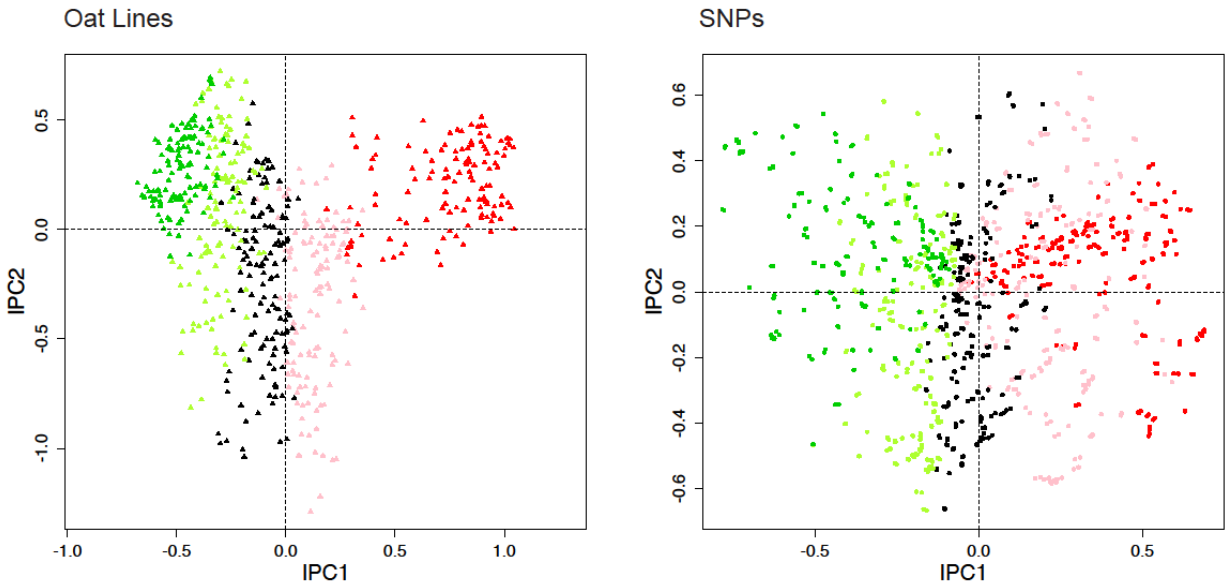

Fig 14. DC-PCA biplot for oat data. The gradient in the CA-arranged matrix in Fig 13 is shown here for both lines and SNPs by the color scheme red, pink, black, light green, dark green.

The DC-PCA graph for oat lines in Fig 14 and the biplot in Fig 12 place each oat line in exactly the same place, so they could be superimposed. Only the color scheme has changed, with colors are assigned by statistical analysis in Fig 14 instead of by geographical information in Fig 12. As would be expected from the diagonal structure in the CA-arranged matrix in Fig 13, the graph for oat lines in Fig 14 shows an arch for the lines, which happens to be an upside-down arch in this case. The gradient in Fig 14 from oat lines marked red at the right to those marked dark green at the left clearly corresponds to the gradient in Fig 12 from Southern US oats marked red at the right to spring oats marked green at the left—and these gradients in both Figs 12 and 14 also correspond to the oat lines in Fig 13 from top to bottom.

The biplot in Fig 14 also shows SNPs on the right. A biplot encourages a unified and even causal interpretation of both lines and SNPs. The SNPs, just like the lines, show an evident gradient from red at the right to dark green at the left. Since the SNP coding rare=1 has been used, this gradient means that the minor allele for SNPs shown in red predominates in Southern

US oats, whereas the minor allele for SNPs shown in dark green predominates in spring oats—and necessarily the opposite holds for the major allele.

The authors of the article that presented our oat example "speculated that adaptation to different planting times and different flowering cues could underlie genetic structure," so they "selected heading date for GWAS as an indicator trait for local adaptation" [18]. Their quantitative trait loci (QTL) scans found strong associations with candidate genes for heading date, *Vrn3* and the linked regulatory gene *CO* which have equivalent genes in other species. Our DC-PCA biplot in Fig 14 could be useful in future research to examine SNPs at the right and left extremes to confirm known genes for local adaptation, as well as to search for new candidate genes.

They also found that the strengths of 23 SNP associations with heading date were "specific to location-year," that is, the particular location and year of each field experiment (their Table 2). Our Fig 13 extends that finding to the overall structure in their SNP data. The dense black at the top left indicates (the minor allele of) relevant SNPs in columns at the left and mostly Southern US oats in rows at the top. At the opposite extreme of the matrix diagonal, the dense black at the bottom right indicates (the minor allele of) relevant SNPs in columns at the right and mostly spring oats in rows at the bottom. Those two groups constitute the most obvious structure in Fig 13, but careful inspection reveals additional interesting structure. About two-thirds of the way to the right (namely SNPs 820–870 in CA1 order), there is a group of about 50 SNPs with rather dense black in the *middle* that are associated with agroecological conditions intermediate between the Southern US and spring extremes. Those SNPs are among the light green points in the biplot in Fig 14, and they are located in a narrow vertical band centered at IPC1 = -0.15. This discovery, that numerous SNPs are concentrated at various positions along the CA1 gradient (top, middle, or bottom) that reflects an agroecological causal gradient, merits further investigation. Of course, many other SNPs or columns in Fig 13 show no structure whatsoever—they are not associated with the contrast in heading date between Southern US and spring oats, although some might be associated with other agroecological causal factors.

Our survey of 125 articles with PCA graphs of SNP data encountered no biplots. This is a lost opportunity. PCA of SNP data using best practices, rather than contemporary practices, can display joint structure with biplots and thereby strengthen inferences causal inferences.

Discussion

On five counts, contemporary PCA analyses of SNP data exhibit shortcomings. (1) The implications of SNP polarity for PCA analyses have not yet been realized, including the advantages of SNP coding rare=1. (2) The best variant DC-PCA has hardly been used, if at all. (3) Crucial choices of SNP coding and PCA variant are reported only rarely in materials and methods. (4) PCA biplots of both Items and SNPs have been absent from the genomics literature, even though biplots could greatly expand the opportunity for SNP data to inform causal inferences. (5) Only about 10% of PCA graphs have axes that are scaled properly.

Wider implications

It has not escaped our attention that the issues noted here may be only the tip of the iceberg: They are likely to arise for other kinds of genomics data and related statistical analyses.

Many other kinds of data besides SNP data are used in genetics, genomics, and breeding. For example, terminal restriction fragment length polymorphism (T-RFLP) data is used extensively to fingerprint microbial communities. Ten representative microbial datasets were used to compare several statistical analyses: T-RFLP-Centered PCA, NMS with Sørensen, Jaccard, and Euclidean distances, CA, detrended correspondence analysis (DCA), and a new method for analysis of T-RFLP data, DC-PCA which they called AMMI [39]. By several criteria, DC-PCA is best. Free, open-source, convenient software is available for DC-PCA analysis of T-RFLP data [40]. However, for many other kinds of genomics data, including gene expression data, the issues raised here have not yet been addressed.

Likewise, many related statistical analyses besides PCA analysis are used in genomics. Several statistical methods are similar to PCA, including principal coordinates analysis (PCoA), MDS, CA, DCA, and t-distributed stochastic neighbor embedding (t-SNE) [41]. Recently t-SNE has attracted much interest [34, 42]. Further evaluation of t-SNE would be prudent, especially since several issues raised here regarding PCA have not yet been considered in the context of t-SNE analysis. PCoA can use the simple matching distance that is invariant to SNP coding and thereby sidesteps the coding issue, but what if any benefits would result has not yet been investigated. PCA using covariance has been compared with NMS using the allele sharing distance metric [43]. DCA has been used extensively in ecology to eliminate the arch distortion of PCA and CA, thereby making secondary gradients more readily visible [44]. Visualization of primary and secondary gradients by biplots that show both species and samples has enabled ecologists to develop rich causal explanations for the structure in their data. DCA has been applied to data on 16S rRNA gene amplicons [45]. The issues raised here are also relevant for other statistical analyses that share the general purpose of displaying data structure, despite being visually quite different from PCA, including dendrograms, heat maps, and neighbor joining trees.

Some statistical methods used widely outside genomics have not yet been applied to genomics data to the best of our knowledge, but might well prove to be effective. For instance, two-way indicator species analysis (TWINSPAN) is used in ecology [46]. It provides integrated dendrograms of both Items and SNPs that can be used to make a heat map which is ideally suited

to reveal joint structure. Comparison of TWINSPAN with the popular Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method and other hierarchical classifications could be interesting. TWINSPAN might prove to be exceptional because it is a polythetic, divisive method based on repeated applications of CA.

Conclusions

Five simple recommendations for effective PCA analysis of SNP data emerge from this investigation.

(1) Use the SNP coding 1 for the rare or minor allele and 0 for the common or major allele.

(2) Use DC-PCA; for any other PCA variant, examine its augmented ANOVA table.

(3) Report which SNP coding and PCA variant were selected, as required by contemporary standards in science for transparency and reproducibility, so that readers can interpret PCA results properly and reproduce PCA analyses reliably.

(4) Produce PCA biplots of both Items and SNPs, rather than merely PCA graphs of only Items, in order to display the joint structure of Items and SNPs and thereby to facilitate causal explanations. Be aware of the arch distortion when interpreting PCA graphs or biplots.

(5) Produce PCA biplots and graphs that have the same scale on every axis.

Software developers play a key role in determining which data analyses are reasonably easy to perform, and which analysis options are the defaults and hence are used most frequently. Greater awareness of best practices for PCA analysis of SNP data presents software developers with new opportunities to enable their user communities to accelerate advances in human genomics and medicine, crop genetics and breeding, and other vital applications.

Appendix: Consequences of SNP polarity for six variants of PCA

This appendix concerns which variants of PCA are, or else are not, immune to changes in SNP coding as regards PCA graphs of Items, where "Items" is a generic term for persons or cultivars or other samples. The main text already showed in Tables 2 and 3 that Item-Centered PCA is not immune because SNP coding affects the SS for SNP main effects, and thereby can change which PC is dominated by the SNP main effects, which can alter dramatically a PCA graph of Items. This same verdict of not being immune also applies to Item-Standardized PCA for the same reason. Likewise, Grand-Mean-Centered PCA is not immune because it also retains SNP main effects (and Item main effects), and again SNP coding affects the SS for SNP main effects. The remainder of this appendix addresses the remaining three variants in the order SNP-Centered, SNP-Standardized, and Double-Centered PCA.

First, consider SNP-Centered PCA. Let Y be the $p \times n$ SNP data matrix with SNPs in p rows and Items in n columns. Without loss of generality, assume that $p \geq n$. The matrix Y may be SNP-Centered as follows: $Y_C = Y\left(I_n - n^{-1}1_n 1_n^T\right)$, where I_n is the n-dimensional identity matrix and 1_n is an n-vector of ones. Let $Y_C = USV^T$ be a singular value decomposition of Y_C, where U is a $p \times n$ orthonormal matrix of left singular vectors holding the row scores, V is an $n \times n$ orthogonal matrix right singular vector holding the column scores, and S is a diagonal matrix of order n holding the ordered singular values. From the orthonormality of U we have $U^T U = I_n$ and from the orthogonality of V we have $V^T V = VV^T = I_n$.

If the polarity of the r-th SNP is changed by swapping 0s and 1s in this r-th row of Y, this operation can be written as $\tilde{Y}_C = PY_C$, where P is a diagonal matrix of order p with $P\{r,r\} = 1$ if the polarity of the r-th SNP is unchanged and $P\{r,r\} = -1$ if the polarity is changed. It is important to note that $PP^T = P^T P = I_p$. Now $\tilde{Y}_C$ can be written as $\tilde{Y}_C = PY_C = PUSV^T = \tilde{U}SV^T$, where $\tilde{U} = PU$. The right-hand side of this equation can be seen to represent an SVD of $\tilde{Y}_C$ because $\tilde{U}^T\tilde{U} = U^T P^T PU = U^T U = I_n$. Thus, V is the matrix of right singular vectors of both Y_C and $\tilde{Y}_C$. For SNP-Centered PCA, this explains why (up to a possible sign change of whole columns) the column or Item scores remain unaltered after changing the polarity of coding (that is, swapping 0s and 1s) for any or all SNPs.

Second, consider SNP-Standardized PCA. For standardized data, $Y_S = D^{-1/2}Y_C$ where $D = diag(W)$ with $W = (n-1)^{-1}Y_C Y_C^T = (n-1)^{-1}Y\left(I_n - n^{-1}1_n 1_n^T\right)Y^T$. Changing the polarity of some SNPs does not change the SNP variances in D. Therefore, the above results for SNP-Centered data carry over fully to SNP-Standardized data.

Third and finally, consider Double-Centered PCA. DC-PCA is not immune to changes in SNP polarity as regards PCA graphs for Items. Double-Centering pertains to the matrix

$Y_{DC} = \left(I_p - p^{-1}1_p 1_p^T\right)Y_C$. If the polarity of some SNPs are changed, then PY_C needs to be computed *before* the centering for Items. Thus, we need to compute $\tilde{Y}_{DC} = \left(I_p - p^{-1}1_p 1_p^T\right)PY_C$. The matrices P and $\left(I_p - p^{-1}1_p 1_p^T\right)$ do not commute; that is, $\left(I_p - p^{-1}1_p 1_p^T\right)P \neq P\left(I_p - p^{-1}1_p 1_p^T\right)$ so $\tilde{Y}_{DC} = \left(I_p - p^{-1}1_p 1_p^T\right)PY_C \neq P\left(I_p - p^{-1}1_p 1_p^T\right)Y_C = PY_{DC}$. Therefore, the SVD of $\tilde{Y}_{DC}$ cannot be obtained from that of Y_{DC} in the same way as the SVD of $\tilde{Y}_C$ can be obtained from that of Y_C. This explains why Item scores before and after changing the polarity of some SNPs are not perfectly correlated.

However, when the sum of squares (SS) for Item main effects is small relative to the SS for SNP-by-Item interaction effects, centering by Item has little effect on the Item scores based on SVD. The verdicts on immunity to SNP coding will be nearly the same for DC-PCA and SNP-Centered PCA when Item main effects are small, and SNP-Centered PCA was already proven earlier in this appendix to be immune. Therefore, correlations for Item scores between different SNP codings are expected to be very close to 1 or -1 for DC-PCA.

Indeed, the main text already gave two examples of different SNP codings, involving comparison of Figs 1 and 3 and comparison of Figs 6 and 7, that have correlations of nearly 1 for IPC1 and IPC2 Item scores. (Recall from the main text that the terminology "interaction principal component" or IPC distinguishes the principal components of DC-PCA from those of all other variants of PCA.) For a third example, consider the oat data as received with arbitrary SNP coding and the re-coded data with rare=1. For arbitrary SNP coding the SS for Item main effects is only 0.34% that for SNP-by-Item interaction effects; and from Table 4 for SNP coding rare=1 only 1.50%. The correlations of IPC1 and IPC2 Item scores between these two datasets are nearly 1 as expected, namely 0.9974405 and 0.9999766. For this oat dataset, again the IPC1-IPC2 graphs of Items are virtually unchanged by differences in SNP coding. A small SS for Item main effects compared to that for SNP-by-Item interaction effects is a necessary and sufficient condition for DC-PCA graphs of Items to be virtually immune to changes in SNP polarity.

Materials and methods

Literature survey

The 125 examples of PCA graphs of SNP data were taken from the literature more or less at random, with some emphasis on recent articles and agricultural crop species. They span many species and many journals. This survey is included in the supporting information (S1 Table).

Constructed and real datasets

All nine datasets used for figures and tables are included in the supporting information (S2 Text).

PCA and CA analyses

Our R code for comparing PCA variants and CA is included in the supporting information (S3 R Code). From the R library, our code uses ca for CA and schoolmath for formatting tables.

Supporting information

S1 Table. Literature survey of 125 articles that apply PCA analysis to SNP data.

S2 Text. All nine datasets used to produce figures or tables.

S3 R Code. R code used to perform PCA and CA analyses.

Acknowledgments

We appreciate helpful comments on this manuscript from Peter Bradbury, Samuel Cartinhour, Kathy Esvelt Klos, and Kelly Robbins. We also appreciate Kathy Esvelt Klos sharing the oat SNP data with us.

Author Contributions

Conceived and designed the inquiry, and invented augmented ANOVA tables and a new method for causal exploration: HG. Analyzed data and visualized figures: SQ. Wrote R software: SQ LZ RC. Conducted literature survey: LZ SQ HG. Wrote the appendix: HP. Wrote the paper: HG. Reviewed the final version of the paper: all authors.

References

1. Gabriel K. The biplot graphic display of matrices with application to principal component analysis. Biometrika. 1971; 58: 453–467. doi: https://doi.org/10.1093/biomet/58.3.453
2. Gower J, Lubbe S, le Roux N. Understanding biplots. New York: John Wiley and Sons; 2011.
3. Malik WA, Piepho H-P. Biplots: Do not stretch them! Crop Sci. 2018; 58: 1-9. doi: https://doi.org/10.2135/cropsci2017.12.0747
4. Chen J, Zheng H, Bei JX, Sun L, Jia WH, Li T, et al. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am J Hum Genet. 2009; 85: 775-785. doi: https://doi.org/10.1016/j.ajhg.2009.10.016 PMID: 19944401
5. Galinsky KJ, Bhatia G, Loh PR, Georgiev S, Mukherjee S, Patterson NJ, et al. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. Am J Hum Genet. 2016; 98: 456-472. doi: https://doi.org/10.1016/j.ajhg.2015.12.022 PMID: 26924531
6. Abraham G, Inouye M. Fast principal component analysis of large-scale genome-wide data. PLoS One. 2014; 9: e93766. doi: https://doi.org/10.1371/journal.pone.0093766 PMID: 24718290
7. Jackson JE. A user's guide to principal components. New York: Wiley-Interscience; 1991.
8. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015; 4: 7. doi: https://doi.org/10.1186/s13742-015-0047-8 PMID: 25722852
9. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27: 2156-2158. doi: https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522
10. Morrison D, Van Iersel L, Kelk S, List M. 2012. Available from: http://phylonetworks.blogspot.com/2012/12/distortions-and-artifacts-in-pca.html.
11. Morrison D, Van Iersel L, Kelk S, List M. 2016 3 May. Available from: http://phylonetworks.blogspot.com/2016/05/continued-misuse-of-pca-in-genomics.html.
12. Gauch HG. Multivariate analysis in community ecology. Cambridge, UK: Cambridge University Press; 1982.
13. Digby PGN, Kempton RA. Multivariate analysis of ecological communities. New York: Chapman and Hall; 1987.
14. Diaconis P, Goel S, Holmes S. Horseshoes in multidimensional scaling and local kernel methods. Ann Appl Stat. 2008; 2: 777-807. doi: https://doi.org/10.1214/08-aoas165
15. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. Nat Genet. 2008; 40: 646-649. doi: https://doi.org/10.1038/ng.139 PMID: 18425127
16. Reich D, Price AL, Patterson N. Principal component analysis of genetic data. Nature Genet 2008; 40: 491-492. doi: https://doi.org/10.1038/ng0508-491 PMID: 18443580
17. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007; 23: 2633-2635. doi: https://doi.org/10.1093/bioinformatics/btm308 PMID: 17586829
18. Esvelt Klos K, Huang YF, Bekele WA, Obert DE, Babiker E, Beattie AD, et al. Population Genomics Related to Adaptation in Elite Oat Germplasm. Plant Genome. 2016; 9. doi: https://doi.org/10.3835/plantgenome2015.10.0103 PMID: 27898836
19. Gauch HG, Piepho H-P, Annicchiarico P. Statistical analysis of yield trials by AMMI and GGE: Further considerations. Crop Sci. 2008; 48: 866-889. doi: https://doi.org.10.2135/cropsci2005.07-0193
20. Pearl J, Mackenzie D. The book of why: The new science of cause and effect. New York: Basic Books; 2018.

21. Dandine-Roulland C, Perdry H. Where is the causal variant? On the advantage of the family design over the case–control design in genetic association studies. Eur J Hum Genet. 2015; 23: 1357-1363. doi: https://doi.org/10.1038/ejhg.2014.284 PMID: 25585700
22. Hong S, Kim Y, Park T. Practical issues in screening and variable selection in genome-wide association analysis. Cancer Inform. 2015; 13: 55-65. doi: https://doi.org/10.4137/CIN.S16350 PMID: 25635166
23. McVean G. A genealogical interpretation of principal components analysis. PLoS Genetics. 2009; 5: e1000686. doi: https://doi.org/10.1371/journal.pgen.1000686 PMID: 19834557
24. Zheng X, Weir BS. Eigenanalysis of SNP data with an identity by descent interpretation. Theor Popul Biol. 2016; 107: 65-76. doi: https://doi.org/10.1016/j.tpb.2015.09.004 PMID: 26482676
25. Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, et al. Reconstructing the population genetic history of the Caribbean. PLoS Genet. 2013; 9: e1003925. doi: https://doi.org/10.1371/journal.pgen.1003925 PMID: 24244192
26. Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, et al. Genomic Insights into the Ancestry and Demographic History of South America. PLoS Genet. 2015; 11: e1005602. doi: https://doi.org/10.1371/journal.pgen.1005602 PMID: 26636962
27. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, et al. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet. 2007; 3: e160. doi: https://doi.org/10.1371/journal.pgen.0030160 PMID: 17892327
28. Chen GB, Lee SH, Zhu ZX, Benyamin B, Robinson MR. EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations. Heredity. 2016; 117: 51-61. doi: https://doi.org/10.1038/hdy.2016.25 PMID: 27142779
29. Caldu-Primo JL, Mastretta-Yanes A, Wegier A, Pinero D. Finding a needle in a haystack: distinguishing mexican maize landraces using a small number of SNPs. Front Genet. 2017; 8: 45. doi: https://doi.org/10.3389/fgene.2017.00045 PMID: 28458682
30. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008; 456: 98-101. doi: https://doi.org/10.1038/nature07331 PMID: 18758442
31. Omberg L, Salit J, Hackett N, Fuller J, Matthew R, Chouchane L, et al. Inferring genome-wide patters of admixture in Qataris using fifty-five ancestral populations. BMC Genetics. 2012; 13: 49. doi: https://doi.org/10.1186/1471-2156-13-49 PMID: 22734698
32. Wang C, Zöllner S, Rosenberg NA. A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS Genet. 2012; 8: e1002886. doi: https://doi.org/10.1371/journal.pgen.1002886 PMID: 22927824
33. Yang J, Jin ZB, Chen J, Huang XF, Li XM, Liang YB, et al. Genetic signatures of high-altitude adaptation in Tibetans. Proc Natl Acad Sci U S A. 2017; 114: 4189-4194. doi: https://doi.org/10.1073/pnas.1617042114 PMID: 28373541
34. Byrne RP, Martiniano R, Cassidy LM, Carrigan M, Hellenthal G, Hardiman O, et al. Insular Celtic population structure and genomic footprints of migration. PLoS Genet. 2018; 14: e1007152. doi: https://doi.org/10.1371/journal.pgen.1007152 PMID: 29370172
35. Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MG. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. Mol Biol Evol. 2016; 33: 1082-1093. doi: https://doi.org/10.1093/molbev/msv334 PMID: 26715629
36. Luu K, Bazin E, Blum MG. pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol Ecol Resour. 2017; 17: 67-77. doi: https://doi.org/10.1111/1755-0998.12592 PMID: 27601374

37. Zeng X, Chakraborty R, King JL, LaRue B, Moura-Neto RS, Budowle B. Selection of highly informative SNP markers for population affiliation of major US populations. Int J Legal Med. 2016; 130: 341-352. doi: https://doi.org/10.1007/s00414-015-1297-9 PMID: 26645290
38. Zeng X, Warshauer DH, King JL, Churchill JD, Chakraborty R, Budowle B. Empirical testing of a 23-AIMs panel of SNPs for ancestry evaluations in four major US populations. Int J Legal Med. 2016; 130: 891-896. doi: https://doi.org/10.1007/s00414-016-1333-4 PMID: 26914801
39. Culman SW, Gauch HG, Blackwood CB, Thies JE. Analysis of T-RFLP data using analysis of variance and ordination methods: a comparative study. J Microbiol Methods. 2008; 75: 55-63. doi: https://doi.org/10.1016/j.mimet.2008.04.011 PMID: 18584903
40. Culman SW, Bukowski R, Gauch HG, Cadillo-Quiroz H, Buckley DH. T-REX: Software for the processing and analysis of T-RFLP data. BMC Bioinformatics. 2009; 10: 171. doi: https://doi.org/10.1186/1471-2105-10-171 PMID: 19500385
41. van der Maaten L. Accelerating t-SNE using tree-based algorithms. J Machine Learning Res. 2014; 15: 3221–3245.
42. Platzer A. Visualization of SNPs with t-SNE. PLoS One. 2013; 8: e56883. doi: https://doi.org/10.1371/journal.pone.0056883 PMID: 23457633
43. Gao X, Martin ER. Using allele sharing distance for detecting human population stratification. Hum Hered. 2009; 68: 182-191. doi: https://doi.org/10.1159/000224638 PMID: 19521100
44. Hill MO, Gauch HG. Detrended correspondence analysis: An improved ordination technique. Vegetatio. 1980; 43: 47-58. doi: https://doi.org/10.1007/BF00048870
45. Zhang X, Qu Y, Ma Q, Zhang Z, Li D, Wang J, et al. Illumina MiSeq sequencing reveals diverse microbial communities of activated sludge systems stimulated by different aromatics for indigo biosynthesis from indole. PLos One. 2015; 10: e0125732. doi: https://doi.org/10.1371/journal.pone.0125732 PMID: 25928424
46. Hill MO, Bunce RGH, Shaw MW. Indicator species analysis, a divisive polythetic method of classification, and its application to a survey of native pinewoods in Scotland. J Ecol. 1975; 63: 597–613. doi: https://doi.org/10.2307/2258738