Consequences of SNP codings, PCA variants, and PCA graphs for elucidating population structure

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 display and discover patterns in SNP data from humans, animals, plants, and microbes—especially to elucidate population structure. Given the popularity of PCA, one might expect that PCA is understood well and applied effectively. However, our literature survey of 125 representative articles that apply PCA to SNP data shows that three choices have usually been made poorly: SNP coding, PCA variant, and PCA graph. Accordingly, we offer several simple recommendations for effective PCA analysis of SNP data. The ultimate benefit from informed and optimal choices of SNP coding, PCA variant, and PCA graph is expected to be discovery of more biology, and thereby acceleration of medical, agricultural, and other vital applications.

Introduction

Single nucleotide polymorphism (SNP) data is common in the genetics and genomics literature, and principal components analysis (PCA) is one of the most frequent statistical analyses applied to SNP data. PCA analyses of SNP data involve three methodological choices concerning SNP codings, PCA variants, and PCA graphs. However, a crucial question has not yet been asked: What are the consequences of these three choices, both singly and in combination, and which choices are better than others for elucidating the structure or patterns in the data? This is a pressing question because of the sheer volume of PCA analyses of SNP data, with applications that span humans, animals, plants, and microbes.

Results

Preliminary matters

Both the Results and the following Discussion address three topics in the same order: SNP codings, PCA variants, and PCA graphs. Because these topics are deeply interrelated, this preliminary material provides just enough background on PCA variants and graphs for the following discussion of SNP codings to make sense. In order to simplify and focus this article, best practices are prioritized from the outset, although unavoidably the principles and arguments that identify best practices develop gradually throughout this article.

The distinction between a PCA monoplot and a PCA biplot is that the former has points for only the Items or only the SNPs, whereas the latter includes both. The generic term used here for either a monoplot or biplot is a "graph." Biplots were first introduced by Gabriel [1] and have become the norm in countless applications of PCA [2], including ecology and agriculture. Biplots are part of best practices, as argued later.

The following principles for interpreting biplots pertain to a specific PCA variant, double-centered PCA (DC-PCA), also part of best practices. As will be further explained later, DC-PCA applies PCA to the matrix of SNP-by-Item (S×I) interactions, and accordingly its principal components are given the special name of interaction principal components (IPCs), which are denoted by IPC1, IPC2 and so on for the first and following IPCs. For two points of the same kind (two Items or else two SNPs), nearby points have similar interactions, whereas distant points have dissimilar interactions. For two points of different kinds (an Item and a SNP), Items in a given direction from the origin have positive interactions for SNPs in that same direction, Items in a given direction have negative interactions for SNPs in the opposite direction, and Items in a given direction have small interactions for SNPs at right angles. In greater detail, S×I interactions are approximated by IPC1 as products of Item and SNP scores, so an IPC1 score of 0.9 for a given Item and 0.5 for a SNP yields an estimated S×I interaction of 0.45. Continuing with IPC2, scores of –0.7 for an Item and 0.4 for a SNP gives an estimated interaction of –0.28; and hence the combined estimate for the IPC1-IPC2 model is 0.45 – 0.28 = 0.17. Accordingly, interactions are large when an Item and SNP are both far from the origin (but not at right angles), whereas an Item near the origin has small interactions with all SNPs, and

likewise a SNP near the origin has small interactions with all Items. More exactly, these interpretive principles address the interaction structure captured in the IPC1-IPC2 plane of a biplot.

By definition, S×I interactions inherently involve *both* SNPs and Items, so a biplot is *required* for PCA to display S×I interactions. By way of preview, although DC-PCA has this single and straightforward set of interpretive principles that applies to any and all datasets, it has this advantage uniquely: All other PCA variants have multiple and complicated sets of interpretive principles that vary from dataset to dataset.

SNP codings

Consider a data matrix comprised of a number of SNPs observed for a number of Items, where "Items" is a generic term for the samples, such as individual humans, horses, cultivars of wheat, or races of a pathogen. 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. We begin with the simplest case of biallelic SNPs.

One option, here called SNP coding rare=1, is to code the rare allele as 1 and the common allele as 0. This coding is of special interest for reasons that emerge momentarily.

Another option, here called SNP coding common=1, is the opposite: to code the common allele as 1 and the rare allele as 0. This coding is the default in TASSEL, which is widely used for crop plants ([3]; Peter Bradbury personal correspondence, 23 April 2018).

The third and final possibility, here called SNP coding mixed, is to code the alleles in some other manner that yields a mixture of rare and common alleles as 1 (and as 0). For example, the variant call format (VCF) distinguishes one Item as the reference genome, and then for each SNP it assigns 0 to the allele of the reference genome and 1 to the other non-reference allelle. VCF is popular because it was developed for the 1000 Genomes Project in human genetics [4], and subsequently has been adopted widely.

The SNP coding rare=1 can be generalized for SNP datasets having more than the 2 codes of biallelic data. If need be, transpose the data to a set of consecutive integer values starting with 0, such as 0, 1, 2, and 3. Then recode each SNP to give the rarest allele the highest value, and so on, and lastly give the most common allele the lowest value of 0. Similarly, for a diploid species with three codes—one for each of the two homozygotes and another for the heterozygote—if need be, transpose the data to 0 and 2 for the homozygotes and 1 for the heterozygote. Then recode each SNP to give the rarest homozygote the value 2 and the most common homozygote the value 0. These ideas can be elaborated for polyploids such as hexaploid (AABBDD) bread wheat, *Triticum aestivum* (L.). The other two codings can also be generalized.

Fig 1 illustrates the 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 ones readily visible. The convention adopted here is to number matrix columns from left to right, and number 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. [5]; also see [6]. The concentration of ones along the matrix diagonal constitutes a single gradient with evident joint structure that involves *both* Items and SNPs.

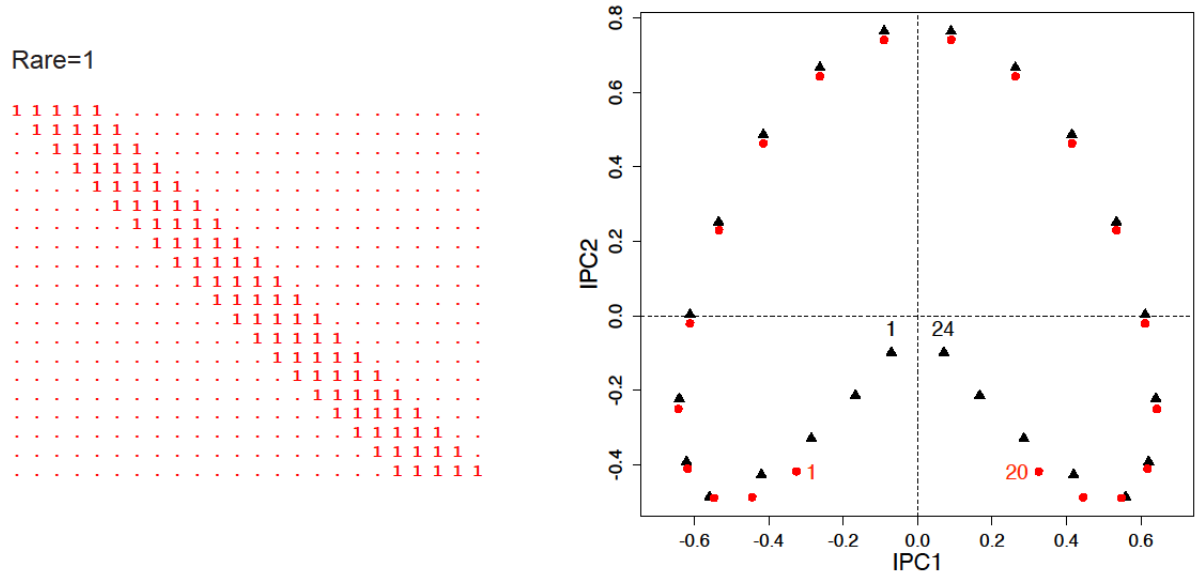

Fig 1. A simple gradient with SNP coding rare=1 and its DC-PCA biplot. The matrix has 24 Items in its columns and 20 SNPs in its rows. The biplot shows Items as black triangles and SNPs as red dots.

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. The 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 so-called arch distortion, the gradient is still apparent: 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 [7, 8]. However, an extensive search of the genomics literature by Morrison found only two papers that discuss the PCA arch distortion [9, 10]. Those papers and his blogposts have not yet succeeded in making this distortion well known (personal correspondence, David Morrison, 18 February 2018).

Fig 2 shows the same gradient as Fig 1, but with the opposite option, SNP coding common=1. The arch for Items 1 to 24 is shown by black triangles, and the arch for SNPs 1 to 20 by green dots. Compared to Fig 1, the arch for SNPs shown in green has been rotated by 180° relative to the Items shown in black. Therefore, Item 1 and SNP 1 are far apart in Fig 2, although they are near each other in Fig 1; and the same applies to Item 24 and SNP 20. On the other hand, Item 17 and SNP 6 nearly touch in Fig 2 (near the horizontal dashed line, at the right), although they are far apart in Fig 1; and the same applies to Item 8 and SNP 15 (near the horizontal dashed line, at the left). Consequently, the biplot in 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.

Common=1

```
. . . . . 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 . . . . . 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 . . . . . 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 . . . . . 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 . . . . . 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 . . . . . 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 . . . . . 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 . . . . . 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 . . . . . 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 . . . . . 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 . . . . . 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 . . . . . 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 . . . . . 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 . . . . . 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 . . . . . 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 . . . . . 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 . . . . . 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 . . . . . 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 . . . . . 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 . . . . .
```

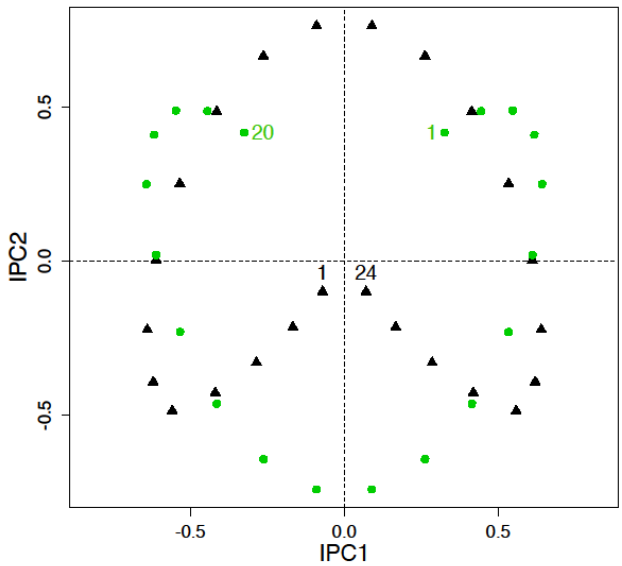

Fig 2. A simple gradient with SNP coding common=1 and its DC-PCA biplot. This is the same matrix as Fig 1 except for the opposite SNP polarity. The biplot shows 24 Items as black triangles and 20 SNPs as green dots.

Fig 3 illustrates SNP coding mixed. 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 instead of every other SNP, or using 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. The orientation of the green arch is rotated by 180° 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. The biplot in Fig 3 inherits the problems from Fig 2 that related Items and SNPs can be widely separated, and distant Items and SNPs can be quite close. Furthermore, Fig 3 has the additional problems that SNPs near each other along the gradient can be far apart in this biplot (such as SNPs 1 and 2), whereas SNPs far apart from each other along the gradient can be near each other (such as SNPs 2 and 13, where 13 is the red dot closest to green dot 2). Consequently, SNP coding mixed produces biplots that are quite confusing.

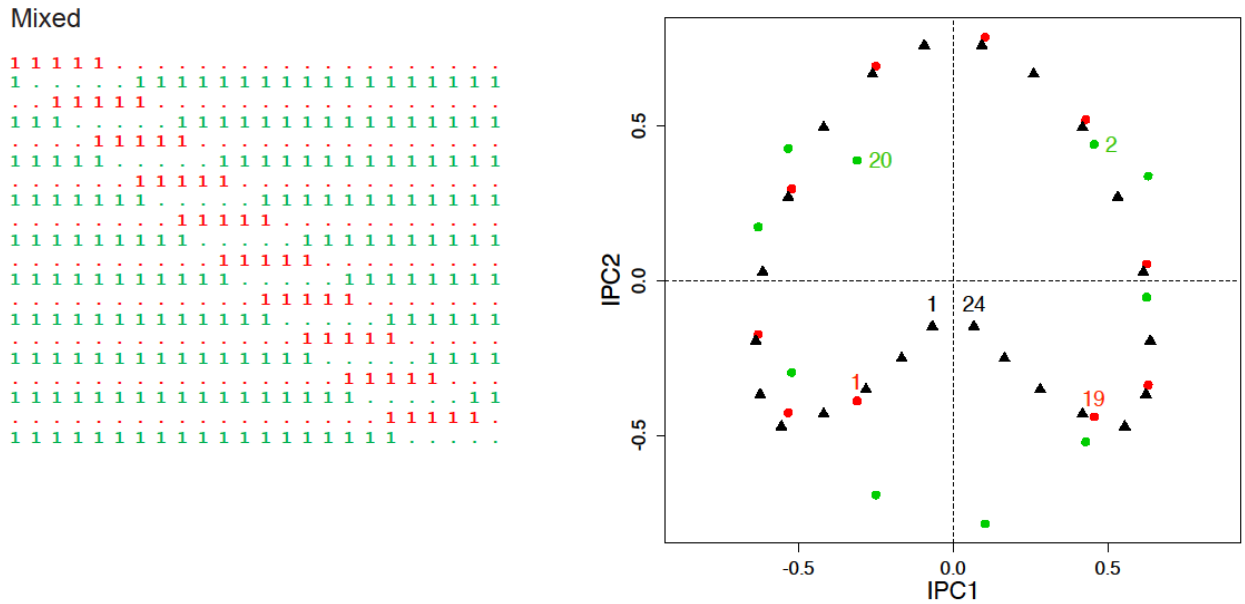

Fig 3. A simple gradient with arbitrary SNP coding and its DC-PCA biplot. This is the same matrix as Figs 1 and 2 except for arbitrary polarity, alternating SNP coding rare=1 (red) and common=1 (green). The biplot shows 24 Items as black triangles, 10 SNPs with coding rare=1 as red dots, and the remaining 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, or both), the results are exactly or nearly identical for the Items (black triangles). Indeed, correlations between Item IPC1 and IPC2 scores for the datasets in Figs 1 and 2 are both exactly 1, and for the datasets in Figs 1 and 3 are 0.9999781 and 0.9983774. Hence, DC-PCA monoplots for Items are virtually immune to differences in SNP coding. The appendix explains this immunity. Nevertheless, as Figs 1 to 3 show, there are huge differences in the patterns for SNPs. However, display of S×I interactions requires *both* Items and SNPs, so non-immunity for SNPs is genuinely problematic.

One more interpretive principle applies to biplots of SNP data, beyond those already explained in the previous subsection on preliminary matters. Whereas Items pertain to one thing, such as individual humans or horses, SNPs pertain to (at least) two things, namely the two polymorphic alleles. If a given SNP were coded twice with rare=1 and again with common=1, it would have two points in a biplot that are located opposite each other relative to the origin, just as Fig 3 would suggest. Consequently, with inconsistent coding, a group of similar SNPs split into *two clusters located opposite each other*.

In review, the three choices of SNP coding, namely rare=1 or common=1 or mixed, are nearly inconsequential for DC-PCA monoplots of only the Items. However, choices of SNP coding are hugely consequential for DC-PCA biplots of both Items and SNPs, for which rare=1

is far superior for displaying structure or patterns in the SNPs in a manner that is readily interpreted. Consequences of SNP coding for PCA variants other than DC-PCA are more complicated, as explained in the next subsection.

PCA variants

A SNPs-by-Items data matrix comprises a two-way factorial design. Analysis of variance (ANOVA) recognizes three sources of variation in such a design: the SNP main effects, Item main effects, and S×I interaction effects [11]. PCA variants emerge from the application of various data transformations prior to PCA analysis. The main three variants discussed here result from subtracting SNP or Item or both effects from the data matrix prior to PCA, and that subtraction is called centering. Hence, these variants are called SNP-Centered, Item-Centered, and Double-Centered PCA. Three additional PCA variants are also mentioned briefly: SNP-Standardized, Item-Standardized, and Grand-Mean-Centered PCA.

The statistical meanings of these three sources of variation are fundamentally different. Consider a data matrix with p SNPs and n Items. The SNP main effects concern the means across Items for each SNP, so they constitute a vector of length p. Likewise, the Item main effects concern the means across SNPs for each Item, so they constitute a vector of length n. By contrast the S×I interactions equal the data minus both main effects, so they constitute a matrix of dimensions p and n. Main effects are relatively simple and can easily be tabulated or graphed by a variety of familiar methods, whereas interaction effects are complex and require multivariate statistical analyses such as PCA. These three sources are radically different statistically in the strong sense that they are orthogonal and uncorrelated, so knowing any one of them provides no information whatsoever on the other two.

The biological meanings of these three sources are as follows in the present context of a SNPs-by-Items data matrix. For a given SNP, its mean across Items is simply the frequency of the allele coded 1 (presuming that the alternative allele is coded 0). Likewise, for a given Item, its mean across SNPs is simply the frequency of the allele coded 1. Ordinarily, it makes sense to avoid burdening or distracting a PCA graph with such simple information, so it is best to remove main effects. Also, these means often lack any straightforward or interesting biological meaning. Indeed, it may be especially difficult to attach any biological meaning to the Item means, not least because of alternative choices for SNP coding. When SNP or Item means are biologically meaningless, they merely add noise to a PCA graph.

By contrast the S×I interactions pertain to the *joint structure* of SNPs-and-Items, that is, the *differential responses or patterns* of Items across SNPs, and of SNPs across Items. These interactions are already familiar to many researchers in other guises, as five examples will show. They concern population structure, SNPs-by-Items heat maps, ancestry-informative SNPs, genome-wide association studies (GWAS), and causal factors. First, interaction is precisely the kind of information that constitutes population structure: different groups of Items associated with different patterns of SNPs, and different groups of SNPs associated with different patterns of Items. Second, when a SNPs-by-Items heat map concentrates one of its colors along the

matrix diagonal, this represents joint structure of SNPs-and-Items, that is, S×I interactions. By contrast, main effects produce horizontal or vertical stripes, rather than diagonal structure. Third, ancestry-informative SNPs that identify specific groups of humans are instances of joint structure of SNPs-and-Items. Fourth, GWAS associates specific SNPs with specific Items that exhibit a given trait (or do so to different degrees), and this joint structure constitutes S×I interactions—*not* SNP main effects, and *not* Item main effects. Fifth and finally, joint structure involving numerous SNPs and numerous Items frequently reflects an underlying causal factor that affects both SNPs and Items. Generalizing beyond these five examples, interests in SNP data are mostly or entirely about their S×I interactions, rather than their main effects.

That said, occasionally there can be special circumstances for which main effects are also of interest, besides the S×I effects, so this situation is also addressed in this article. Which of these three sources of variation—SNP, Item, and S×I effects—are of interest depends on the data and the research objectives. It is the prerogative of researchers to decide which of these sources interest them, so we respectfully leave those choices of interests to others.

The choice among PCA variants can be explored with a recent study concerning environmental adaptation of oats (*Avena sativa* L.) [12]. Kathy Esvelt Klos kindly shared with us a dataset with 635 lines of oats by 1341 SNPs ([12]; 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. We transposed her data to 0 and 1 for the sake of convenience, and we refer to this as the "received data." It had mixed SNP coding. To obtain SNP coding rare=1, polarity was reversed for 772 of the 1341 SNPs. The oat data are included in these two codings in the supporting information (S1_OatMixed and S2_OatRare1).

Table 1 shows the ANOVA table for DC-PCA of the oat data using three SNP codings: rare=1, VCF with oat line 189 as the reference genome (which is of particular interest because it has a larger SS for Item main effects than any of the other 634 possibilities), and the SNP coding for the data as received (and transposed). ANOVA partitions the total degrees of freedom (df) and sum of squares (SS) into three sources, and then PCA partitions the S×I interactions into the first seven IPCs followed by the residual. Indenting of the sources indicates subtotals.

Table 1. ANOVA table for DC-PCA of SNP data on oats using three SNP codings: rare=1 (and common=1 is identical), VCF with oat line 189 as the reference genome, and the data as received. Both VCF189 and the received data are instances of SNP coding mixed.

Source	df	SS rare=1	SS VCF189	SS received
Total	851534	157442.756	175776.962	210084.641
SNPs	1340	16872.731	35206.937	69514.616
Items	634	2077.808	6286.978	482.151
SxI	849560	138492.217	134283.047	140087.874
IPC1	1973	15566.723	12202.771	16158.469
IPC2	1971	9512.804	9172.430	9687.670
IPC3	1969	5836.915	6163.309	6246.232
IPC4	1967	4831.287	4140.115	4803.660
IPC5	1965	3507.237	3478.698	3488.106
IPC6	1963	2997.351	3100.062	3181.724
IPC7	1961	2887.575	2839.061	2881.938
Residual	835791	93352.327	93186.601	93640.076

For SNP coding rare=1, the total SS is composed of 88.0% for S×I interaction effects, 10.7% for SNP main effects, and 1.3% for Item main effects. SNP coding common=1 is not shown in Table 1, but it necessarily has exactly the same ANOVA table as SNP coding rare=1, not only for DC-PCA shown here, but also for all six variants of PCA considered in this article. However, VCF for oat 189 has different percentages, namely 76.4%, 20.0%, and 3.6%, and the received data has 66.7%, 33.1%, and 0.2%. Hence, choices of SNP coding affect the relative magnitudes of these three sources, as well as the relative magnitudes of the IPCs.

The application of PCA to a combination of two sources of variation, unlike the single source for DC-PCA in Table 1, requires a new approach in order to understand what kind of information is in each PC, namely an augmented ANOVA table that is introduced here for the first time. The SS of each PC is partitioned into the portions due to main and interaction effects. 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. Because augmented ANOVA tables are not yet included in

available software, we wrote our own R code, which is included in the supporting information (S3_Software).

Table 2 shows an augmented ANOVA table for SNP-Centered PCA of the same oat data, using SNP coding rare=1. This variant of PCA removes only SNP main effects, and then applies PCA to the Item main effects *and* S×I interaction effects combined, denoted by I&S×I, which has a SS of 2077.808 + 138492.217 = 140570.025. Researchers who are familiar with PCA and are accustomed to the automatic monotonic decrease in the SSs for successive PCs should note that the SSs for the Items and S×I portions are not necessarily monotonic.

Table 2. Augmented ANOVA table for SNP-Centered PCA of SNP data on oats, using SNP coding rare=1. 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	157442.756		
SNPs	1340	16872.731		
I&SxI	850194	140570.025	2077.808	138492.217
PC1	1974	16325.612	857.676	15467.936
PC2	1972	9751.158	257.217	9493.941
PC3	1970	6250.504	423.134	5827.370
PC4	1968	4860.910	23.386	4837.524
PC5	1966	3510.312	4.521	3505.791
PC6	1964	3181.953	187.615	2994.338
PC7	1962	2891.656	1.654	2890.002
Residual	836418	93797.920	322.606	93475.314

All seven PCs and the residual of SNP-Centered PCA contain a mixture of Item and S×I effects. Such mixtures always occur for any dataset whenever PCA is applied to a combination of main and interaction effects [13]. For this particular dataset, the first seven PCs and the residual are all dominated by S×I interaction effects because the Item main effects happen to be small. That outcome could be expected from Table 1 since IPC1 through IPC7 are all larger than the Item main effects.

A great advantage of DC-PCA is that its biplots have a straightforward interpretation, explained earlier, because every IPC contains only one kind of information, S×I interactions. Unfortunately, PCA analyses other than DC-PCA are more difficult to interpret because every PC contains two or three kinds of information.

SNP-Centered PCA has four possible outcomes. The oat example in Table 2 illustrates one possibility, that S×I information dominates both PC1 and PC2. Indeed, the Item main effects account for only 4.3% of the SS captured in a PC1-PC2 graph. Another possible outcome, caused by main effects having a larger SS than does PC1, is that PC1 contains mostly main-effect information whereas PC2 contains mostly S×I information. Yet another possible outcome, caused by main effects having a larger SS than PC2 but a smaller SS than PC1, is the reverse, that PC1 contains mostly S×I information and PC2 contains mostly main-effect information. Finally, it is also possible for a PC to contain substantial portions of both main and interaction effects. For example, from Table 1, SNP-Centered PCA using VCF coding with oat line 189 as the reference genome has a PC1 comprised of 31.1% Item main effects and 68.9% S×I interaction effects. It is crucial for researchers to know which of these four cases obtains for a given dataset when they interpret a PC1-PC2 graph that uses SNP-Centered PCA.

Item-Centered PCA also has four possible outcomes. Table 3 shows the augmented ANOVA table for Item-Centered PCA of the same oat data, using SNP coding rare=1. This variant of PCA removes Item main effects and then applies PCA to the SNP main effects and S×I interaction effects combined, denoted by S&S×I, which has a SS of 16872.732 + 138492.217 = 155364.949. The table shows that PC1 is dominated by SNP main effects (96.2%), whereas PC2 is dominated by S×I interaction effects (99.9%). That outcome could be expected from Table 1 since the SNP main effects are larger than IPC1. As already explained for SNP-Centered PCA in Table 2, this example illustrates only one of the four possible outcomes.

Table 3. Augmented ANOVA table for Item-Centered PCA of SNP data on oats, using 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	851534	157442.756		
Items	634	2077.808		
S&SxI	850900	155364.948	16872.731	138492.217
PC1	1974	17402.420	16734.514	667.906
PC2	1972	15564.018	21.919	15542.099
PC3	1970	9476.489	40.283	9436.206
PC4	1968	5781.809	23.612	5758.197
PC5	1966	4758.747	25.746	4733.001
PC6	1964	3482.969	5.452	3477.517
PC7	1962	2970.194	4.894	2965.270
Residual	837124	95928.303	16.312	95912.021

Like their centered counterparts, PCs from SNP-Standardized and Item-Standardized PCA contain a mixture of main and interaction effects, so these PCA variants also have four possible outcomes. Furthermore, SNP-Standardized PCA, unlike SNP-Centered PCA, cannot be used with VCF coding because the reference genome has a standard deviation of zero.

Finally, Grand-Mean-Centered PCA produces a mixture of SNP main effects, Item main effects, and S×I interaction effects in each PC. Therefore, the situation for Grand-Mean-Centered PCA is quite complex and undesirable: It has seven possibilities, not counting additional possibilities involving a PC with a substantial mixture of main and interaction effects. The supporting information includes the augmented ANOVA tables for these additional variants, using the oat data with SNP coding rare=1 (S4_ThreeTables).

In review, both SNP coding and PCA variant can affect which kind of information—Item or SNP or S×I effects—dominates in each PC. An augmented ANOVA table quantifies the outcome for any dataset, and thereby facilitates proper interpretation of PCA results and graphs. DC-PCA is unique among PCA variants in that it applies PCA to a single source of variation, namely S×I interactions, so its IPCs always contain this one kind of information and hence there is no need for its ANOVA table to be augmented.

PCA graphs

Applications of PCA to SNP data usually include graphs to display patterns or structure in the data. Most commonly, they show PC1 and PC2. When main effects are not of interest, we recommend using DC-PCA to make PCA graphs for the reasons given above. When they are, we recommend the Additive Main effects and Multiplicative Interaction (AMMI) model, introduced below.

Fig 4 shows the biplot for DC-PCA analysis of the oat data, using SNP coding rare=1. To reduce clutter, the biplot is shown in two panels, with oat lines on the left and SNPs on the right. From Table 1, IPC1 and IPC2 capture 11.2% and 6.9% of the S×I interactions, for a total of 18.1%. Experienced oat breeders had classified the 635 oat lines into three groups: 411 spring oats shown in green, 103 world diversity oats in blue, and 121 winter oats, which are also called Southern US oats, in red. We then classified the 1341 SNPs into three corresponding groups, again using the same color scheme. Each SNP assignment is based on which of the three oat groups has the highest percentage of the rare allele, and the outcomes were: 372 highest percentage for spring oats shown in green, 345 highest for world diversity oats in blue, and 624 highest for winter oats in red. The spring (green) and winter (red) oats are expected to cluster and to contrast, whereas the world diversity oats (blue) are heterogeneous and are not expected to cluster.

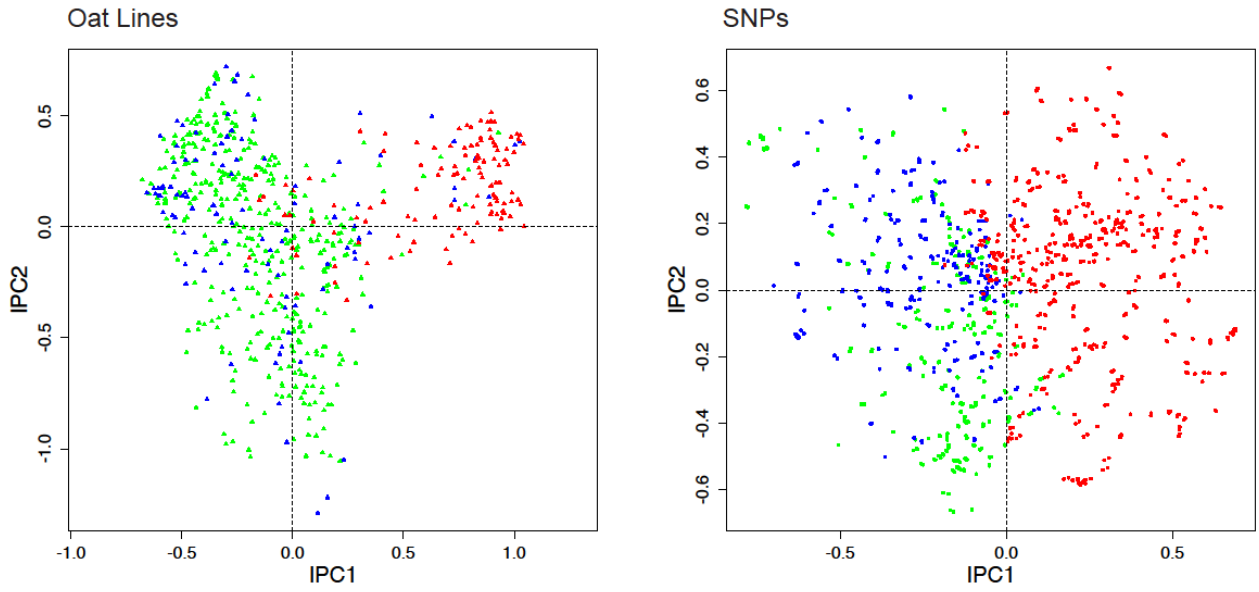

Fig 4. The DC-PCA biplot for the oat data, using SNP coding rare=1. To reduce clutter the biplot uses two panels, with oat lines on the left and SNPs on the right. The 635 oat lines are classified in three groups: 411 spring oats shown in green, 103 world diversity oats in blue, and 121 winter oats in red. Likewise, the 1341 SNPs are classified in three groups based on which oat group has the highest percentage of the rare allele: 372 highest in spring oats shown in green, 345 highest in world diversity oats in blue, and 624 highest in winter oats in red.

The joint structure involving S×I interaction information is quite obvious for IPC1, with green mostly on the left and red mostly on the right *in both panels*. Hence, the 121 winter oats (red) have positive interactions with the 634 SNPs colored red, and the 411 spring oats (green) with the 372 SNPs colored green. The opposite also applies, that the winter oats (red) have negative interactions with the SNPs colored green, and the spring oats (green) with the SNPs colored red. Interactions are large (either positive or negative) for oats and SNPs that are both located near the left or right extremes, but interactions are small for oats or SNPs located near the middle where IPC1 = 0. The overall pattern for the oat lines shows the common arch distortion, in this case an upside-down arch. IPC2 does not provide any additional separation between points of different colors for either oats or SNPs. Other approaches besides the simple method used for Fig 4 could be used to color the Items and SNPs in order to highlight S×I interaction structure, based on statistical analyses or biological information.

When main effects are of interest, we recommend the AMMI model, which combines ANOVA for the main effects with PCA for the multiplicative effects [13]. AMMI and DC-PCA are similar and have an identical ANOVA table. The salient difference is that whereas DC-PCA discards the main effects, AMMI retains them. An AMMI1 biplot shows both of the main effects in its abscissa, and IPC1 in its ordinate; it can show only one component (and hence the suffix 1 in AMMI1) because the main effects use one of its two axes. Its abscissa captures 100% of both main effects. Also, its ordinate captures as much of the S×I interaction effects as possible because IPC1 is the unique least-squares solution that maximizes the variation along this axis and minimizes the residual variation off this axis. We have not yet encountered AMMI in genomics, but it is commonplace in the literature on agricultural yield trials [13]. Incidentally, in statistical analyses of agricultural yield trials, the so-called AMMI2 biplot shows IPC1 and IPC2, which is exactly what a DC-PCA biplot (ordinarily) shows, so “AMMI2” and “DC-PCA” are two names used in different literatures for the same thing.

The interpretive principles for an AMMI1 biplot are that displacements along the abscissa reflect differences in main effects, whereas displacements along the ordinate reflect differences in S×I interaction effects. An Item and SNP with IPC1 scores of the same sign have a positive S×I interaction, whereas those with opposite signs have a negative S×I interaction, and an Item (or SNP) with a score near zero has small interactions with all SNPs (or Items)—at least for those interactions that are captured in the AMMI1 biplot. Like DC-PCA, because AMMI1 has a single kind of information in its abscissa and in its ordinate, an AMMI1 biplot has a single set of interpretive principles that applies to all datasets.

Fig 5 shows the AMMI1 biplot for the oat data, using SNP coding rare=1, the same color scheme as Fig 4, and two adjacent panels. The abscissa shows the mean frequency of the rare allele. The oat lines have means that range from 0.15958 to 0.44893, and the SNPs range from 0.01732 to 0.49921. The vertical line is located at the grand mean of 0.24484. The abscissa captures 100% of the SNP main effects, 100% of the oat line main effects, and 0% of the S×I interaction effects. The ordinate of Fig 5 shows IPC1, which captures 11.2% of S×I, and this ordinate is identical to the abscissa in Fig 4; as before IPC1 separates red from green and blue for

both oat lines and SNPs. In the left panel, the Items have a slight slant because the green and blue group has more oat lines than the red group: Often the larger group determines the common allele which is coded 0, whereas the smaller group tends to have the rare allele coded 1. In the right panel, the SNPs have a broad range of IPC1 scores at the right but not at the left because SNPs at the right have large numbers of both rare and common alleles and hence have large variances, whereas SNPs at the left have mostly the common allele and hence have small variances.

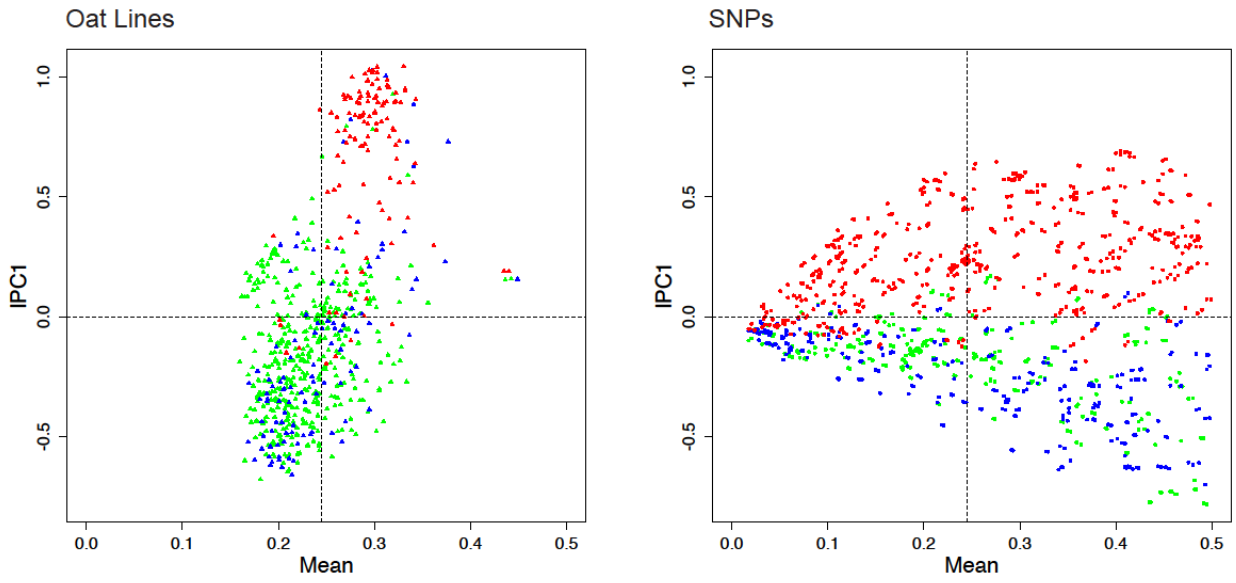

Fig 5. The AMMI1 biplot for the oat data, using SNP coding rare=1. To reduce clutter the biplot uses two panels, with oat lines on the left and SNPs on the right. The color scheme is the same as in Fig 4, namely spring oats show in green, world diversity oats in blue, and winter oats in red, with corresponding colors for the SNPs.

Five additional PCA biplots for the oat dataset are shown in the supporting information, using SNP coding rare=1 and the same color scheme as the main text (S5_FiveBiplots). SNP-Centered and SNP-Standardized PCA approximate DC-PCA in Fig 4 because the Item main effects are small, whereas Item-Centered, Item-Standardized, and Grand-Mean-Centered PCA approximate AMMI1 in Fig 5 because IPC1 captures mostly the large SNP main effects. However, although these other PCA variants can approximate DC-PCA or AMMI1, they can *only* approximate because every component has a mixture of main and interaction effects.

In review, S×I interaction information is displayed by any PCA (or AMMI) biplot, whereas it is completely absent in any PCA monoplot of only Items or only SNPs. An AMMI1 biplot displays main and interaction effects without confounding them.

Discussion

Having characterized the mathematical and statistical consequences of choices of SNP codings, PCA variants, and PCA graphs, here we recommend best practices and contrast them with contemporary practices by means of a literature survey.

Merits of SNP coding rare=1

Based on the results in the previous section, we recommend SNP coding rare=1 for two reasons.

First, regarding the SNP panel of a biplot (or equivalently a SNP monoplot), recall from Figs 1 to 3 that SNP coding rare=1 provides the most sensible and interpretable biplot, with related SNPs and Items located in the same direction relative to the origin. By contrast, SNP coding common=1 places SNPs opposite related Items, making biplots hard to interpret. Likewise, SNP coding mixed (such as VCF) also confuses relationships between SNPs and Items; and worse yet, it confuses relationships among SNPs and among Items, making biplots very hard to interpret.

Second, regarding both the SNP and Item panels of a biplot, recall from Tables 1 to 3 that for any PCA variant other than DC-PCA, the relative magnitudes of the SSs for SNP main effects, Item main effects, PC1, and PC2 can change which kinds of information (main effects, or interaction effects, or a substantial mixture of both) are captured in PC1 and PC2. Both PCA variant and SNP coding can affect those relative magnitudes, and thereby alter PCA graphs drastically—as well as complexify the interpretive principles for these graphs considerably.

Fig 6 further explores VCF coding of SNPs by showing the consequences of selecting each of the 635 oat lines to be the reference genome and then plotting the SS for Item main effects on the abscissa and the SS for SNP main effects on the ordinate. The color codes are the same as before: 411 spring oats shown in green, 103 world diversity oats in blue, and 121 winter oats in red. The centroid for these points is shown by the black plus sign (+) near the middle. For comparison, two additional SNP codings are shown. The orange point at the top left shows the SS for Item main effects and the SS for SNP main effects for the SNP data as received. The brown point at the bottom represents SNP coding rare=1 (or alternatively common=1).

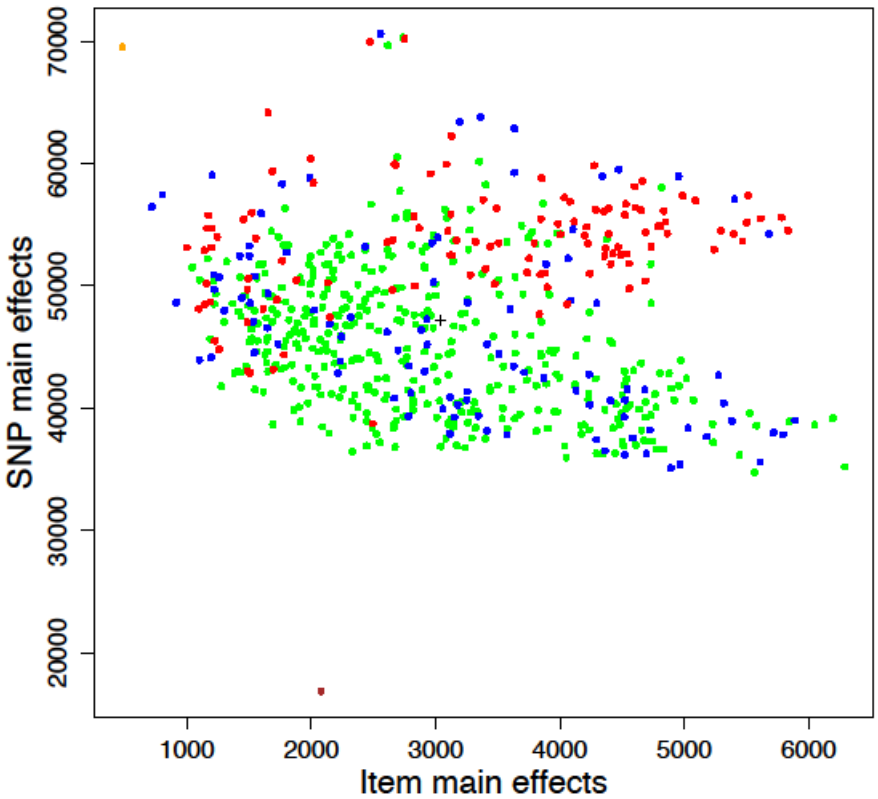

Fig 6. The SS for Item main effects and for SNP main effects, using VCF coding with each of the 635 oat lines selected as the reference genome. These oat lines are coded green for the 411 spring oats, blue for the 103 world diversity oats, and red for the 121 winter oats. The additional orange point near the top left shows results for the received data with SNP coding mixed, and the brown point at the bottom represents SNP coding rare=1 and common=1.

To characterize contemporary practices, we conducted a literature survey of 125 articles that apply PCA to SNP data. This survey is included in the supporting information (S6_LiteratureSurvey). Apparently the popularity of different codings varies among communities: Researchers with human SNP data often use VCF coding because of the influence of the 1000 Genomes Project, whereas researchers with crop SNP data often use SNP coding common=1 because it is the default for TASSEL. We did not notice any unambiguous specification of SNP coding rare=1. However, beyond these broad observations, we cannot provide quantitative results about contemporary practices because the choice of SNP coding is reported so infrequently.

Merits of DC-PCA and AMMI1

We recommend DC-PCA for analyzing SNP data, rather than any of the other five variants considered here, for the following reasons.

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 be removed. Since SNP main effects have already been removed from SNP-Centered PCA, the result is 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 in Table 2—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, then an AMMI1 biplot is a better alternative because it includes both main and interaction effects, but without confounding them. In no case is there any reason to choose SNP-Centered PCA.

The argument against Item-Centered PCA has exactly the same form as the argument against SNP-Centered PCA.

The argument against the last three PCA variants, namely SNP-Standardized, Item-Standardized, and Grand-Mean-Centered PCA, is that all of them fail to resolve the underlying problem of intermixed main and interaction effects. For example, consider the augmented ANOVA analyses in Tables 2 and 3 and in the supporting information (S4_ThreeTables). The proportion of main effects is 1.5% for SNP-Centered PCA and 1.7% for SNP-Standardized PCA, with the remainder being interaction effects. Likewise, the proportion is 10.9% for Item-Centered PCA and 11.0% for Item-Standardized PCA. For Grand-Mean-Centered PCA, the proportions are 1.3% for Item main effects, 10.7% for SNP main effects, and 88.0% for S×I interaction effects.

We strongly recommend that researchers who use any variant of PCA other than DC-PCA should examine 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 that the PCA variant is specified quite infrequently. We did not encounter any specification of DC-PCA, so its applications to SNP data must be quite rare. On the other hand, the great majority of articles do specify the software used for PCA analysis, and no less than 15 different software options were encountered. Software selection may influence the choice of PCA variant, either knowingly or unknowingly. For example, TASSEL has a default of SNP-Centered PCA and an option of SNP-Standardized PCA ([3]; Peter Bradbury personal correspondence, 23 April 2018).

Merits of biplots

We recommend PCA biplots, rather than monoplots. It is axiomatic that a biplot with points for both Items and SNPs is more informative than a monoplot with points for only Items or only SNPs.

Furthermore, *only* a biplot can display the important joint structure of the SNPs-and-Items, that is, the S×I interactions. This is crucial because the recommended DC-PCA analysis shows S×I interactions exclusively. For any other PCA variant, if much or most of the variation in the PCA analysis is S×I interaction—especially if PC1 or PC2 (or both) is nearly all

interaction—then a PCA monoplot is manifestly suboptimal because it *cannot* display interactions.

That S×I interactions inherently involve both SNPs-and-Items, and hence require biplots, can be illuminated by a concrete example. Fig 4 can be interpreted as a monoplot by considering only the left panel, or else only the right panel. The monoplot in the left panel is what the contemporary literature typically features. What does it show? The spring oats (green) are clustered, as are the winter oats (red), whereas the world diversity oats (blue) are dispersed as expected. Occasional exceptions may be of interest, namely a few green points in the mostly red region, and the reverse. But what does it not show? This monoplot for oat lines does not show anything about the genetic differences and SNPs associated with the contrast between spring and winter oats, so it cannot convey any information about S×I interactions. The monoplot for SNPs in the right panel is rarely (if ever) provided in the contemporary literature. Although a SNP monoplot can show clusters and contrasts among SNPs, it also cannot convey any information about S×I interactions. Only when Fig 4 is regarded as a biplot can the interrelated SNPs-and-Items patterns *in both panels* reveal S×I interactions.

The display of S×I interactions by biplots is crucial because S×I is commonly large. From Table 1, the oat data contains 88.0% interaction information (138492.217 / 157442.756), using SNP coding rare=1. Several similar examples can be cited for other crop species, using whatever SNP coding the original authors selected: rice (*Oryza sativa* L.) [14] has 91.5%, soybean (*Glycine max* (L.) Merr.) [15] has 88.7%, maize (*Zea mays* L.) [16] has 65.2%, and potato (*Solanum tuberosum* L.) [17] has 30.1%. Furthermore, the percentage of interaction information captured in a PC1-PC2 graph is often higher than that in a dataset as a whole, as quantified by an augmented ANOVA table. For instance, the percentages of interaction in a PC1-PC2 biplot for SNP-Centered PCA for these five species in the same order are 95.72%, 98.34%, 94.49%, 73.73%, and 99.93%.

The objection may be raised that PCA biplots would be impractical for datasets with many thousands of SNPs, making graphs unworkably cluttered. In fact, high-density PCA graphs appear in the literature routinely, such as Fig 4 in [18] showing results for 54734 humans. Producing biplots in two adjacent panels helps to reduce clutter by separating Items from SNPs. Fortunately, the literature offers several strategies for simplifying PCA graphs. One possibility is to reduce the number of SNPs prior to PCA, using tools such as PLINK [19] and bigstatsr or bigsnpr [20]. Another is to select SNPs of particular interest. For example, only 23 SNPs out of over 1,000,000 produce PCA graphs with clear clusters for several major US populations [21, 22]. Obviously, it is impossible to label thousands of points without causing severe overprinting, but when only a moderate number of Items or SNPs are of special interest, they can be labeled.

In our literature survey, we did not encounter even one biplot.

The winning combination

The contemporary literature on PCA analysis of SNP data exhibits a multiplicity of options for the choices of SNP codings, PCA variants, and PCA graphs. The great multiplicity of

combinations of these choices imposes considerable complexity on this article. By contrast, our recommendations for best practices can be expressed in one concise sentence: Choose SNP coding rare=1, PCA variant DC-PCA (or AMMI1), and biplots.

Even a single departure from that winning combination can obliterate or preclude display of S×I interactions, that is, display of the joint structure of SNPs-and-Items.

For instance, Fig 7 is like Fig 4 in using DC-PCA to produce a biplot, but it departs from recommended practices in just one way: the SNP coding is mixed, rather than rare=1. Although the pattern for the oat lines or Items in the left panel of Fig 7 is nearly identical to that in Fig 4, the pattern for the SNPs in the right panel is utterly obliterated. The explanation is that with the mixed coding of the received data, about half of the SNPs (772 out of 1341) have the reverse polarity common=1, which sends their points to the opposite location in the DC-PCA biplot, and thereby thoroughly mixes the three colors of points for the SNPs. This outcome could be anticipated from Fig 3. Crucially, to display and understand the joint structure of SNPs-and-Items, *both* panels of a biplot must show clear patterns.

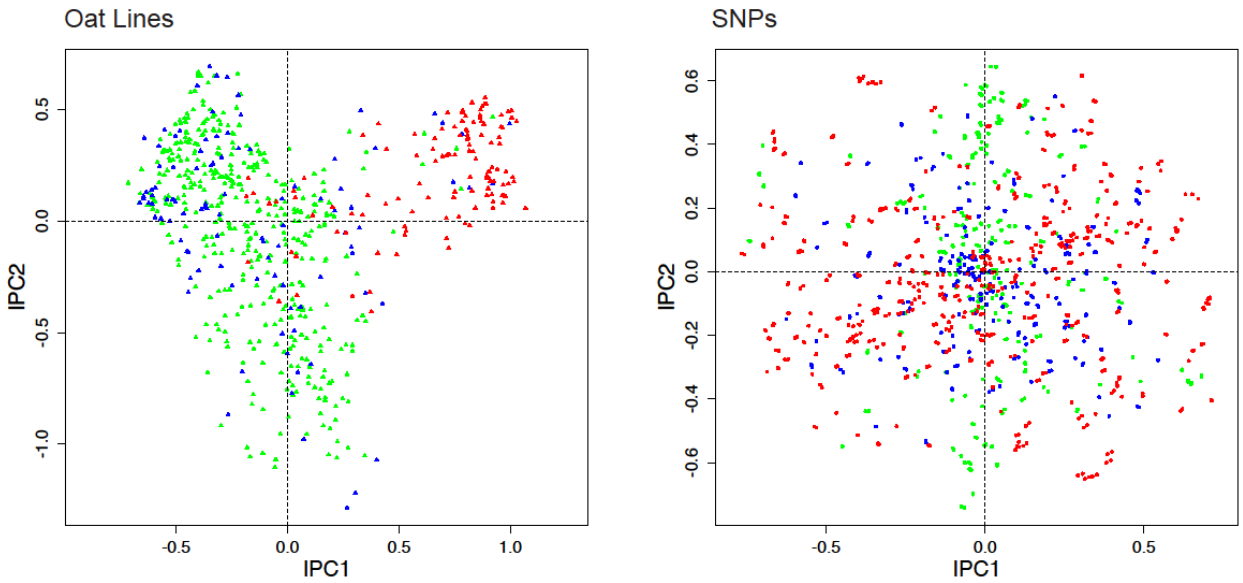

Fig 7. The DC-PCA biplot for the oat data, using SNP coding mixed, namely the received data. The color scheme is the same as in Figs 4 and 5, namely spring oats show in green, world diversity oats in blue, and winter oats in red, with corresponding colors for the SNPs. The contrast between spring and winter oats is clear for Items in the left panel, but the corresponding pattern for SNPs in the right panel has been obliterated by thorough intermixing of the three colors of points.

Likewise, choosing a PCA variant other than DC-PCA is problematic, even if SNP coding rare=1 and a biplot are used as recommended. Examples are shown in the supporting information (S5_FiveBiplots). Briefly, these other variants intermix main and interaction effects in every PC, which complicates the interpretation of PCA graphs.

Finally, regardless whether the recommended SNP coding rare=1 and PCA variant DC-PCA are chosen, the one choice of a monoplot rather than a biplot suffices to eliminate any possibility of displaying and understanding S×I interactions.

In review, our literature survey encountered no clear implementation of *even one* of our three recommendations—SNP coding rare=1, PCA variant DC-PCA, and biplots. Consequently, the likelihood that any published PCA analysis of SNP data has yet implemented *all three* recommendations is quite small.

Conclusions

Necessarily and unavoidably, every PCA analysis of SNP data implements choices of SNP coding and PCA variant. Unfortunately, our survey of 125 articles in which PCA is applied to SNP data shows that these choices are rarely reported, explained, or justified.

Because PCA monoplots of only Items provide some insight into population structure, they are deservedly popular in the literature. This article can explain that success. Because only Items are graphed, unknowingly but luckily the choice of SNP coding—which is often unreported—is nearly inconsequential. And because the SS for Item main effects is often small relative to the SS for IPC1, unknowingly but luckily some other PCA variants, namely SNP-Centered and SNP-Standardized PCA, may approximate the recommended DC-PCA. However, this success has serious limitations. The obvious problem is that relying on luck to obtain a usable monoplot of Items is not as reliable as deliberately choosing the recommended SNP coding rare=1 and PCA variant DC-PCA. The main problem is that a monoplot cannot show interaction structure, which is often the dominant source of variation in a dataset and is usually the variation of principal interest. Production of a useful biplot is an unlikely prospect apart from understanding the consequences of SNP codings and PCA variants.

Three principal recommendations emerge from this investigation into PCA analysis of SNP data. (1) Use the SNP coding 1 for the rare allele and 0 for the common allele. (2) Use the PCA variant DC-PCA if only S×I interactions are of interest, as is often the case; otherwise, use AMMI1 if main effects are also of interest. (3) Use biplots, not monoplots, since only they can display interaction information. Additionally, report which SNP coding and PCA variant were selected, and ideally also provide reasons for those particular choices, so that readers can interpret PCA results properly and reproduce PCA analyses reliably. Finally, if the recommended DC-PCA (or AMMI1) is not used, then provide an augmented ANOVA table in order to quantify the amount of main and interaction effects in each PC.

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. We encourage them to consider including SNP coding rare=1, PCA variant DC-PCA, and biplots for DC-PCA and AMMI1 among the readily available options.

How much important structure has been present in SNP data ever since it was collected, but cannot be displayed or understood by contemporary practices? More specifically, have S×I

interactions escaped attention even though they are biologically interesting? These open questions merit further investigation.

Materials and methods

Construction of PCA graphs

As explained further in the following appendix, each PC is associated with a unit eigenvector for SNPs, a unit eigenvector for Items, and a singular value. For a comprehensive explanation of ways to handle the singular value, see Malik and Piepho [23]. Two brief remarks suffice for present purposes. First the PC scores that appear in this article's biplots were obtained by multiplying both eigenvectors by the square root of the singular value, as is commonly done for biplots. Second, other contexts may call for a different approach, such as a PCA monoplot of Items, for which multiplication by the singular value is preferable because this optimizes the two-dimensional approximation of distances between Items.

The axes in PCA graphs are often scaled to obtain a convenient shape, but actually the axes should have the same scale for many reasons [23]. Unfortunately, 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. Also, 7% of the articles failed to show the scale on one or both PCA axes. However, the two axes of an AMMI1 biplot contain different kinds of information (main or interaction effects), so they do *not* need to use the same scale.

PCA gives a unique least-squares solution, up to simultaneous sign change of the Item and SNP scores for a given PC. Therefore, different software packages applied to the same data may produce PCs with reverse polarity, but that is mathematically inconsequential.

The percentage of variation captured by each PC is often included in the axis labels of PCA graphs. 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 and 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 [24]). 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 the choice of a PCA variant. Readers cannot meaningfully interpret the percentages of variation captured by PCA axes when authors fail to communicate which variant of PCA was used.

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 [18]. The software flashpca is also very fast [25]. The power method is the simplest algorithm for PCA and is efficient when only the first few PCs are needed [26].

Literature survey

The 125 articles applying PCA analysis to SNP data were taken from the literature more or less at random, with some emphasis on agricultural crop species and on researchers at Cornell University. They span many species and many journals. This survey is included in the supporting information (S6_LiteratureSurvey).

Oat datasets

The oat dataset supplied by Kathy Esvelt Klos is included in two formats: SNP coding mixed is the data as received, except that the original coding of 1 and 2 was transposed to 0 and 1; and SNP coding rare=1, which required polarity reversal for 772 of the 1341 SNPs (S1_OatMixed and S2_OatRare1).

PCA and CA analyses

Our R code for comparing six PCA variants and correspondence analysis (CA) is included in the supporting information (S3_Software). From the R library, our code uses ca for CA, schoolmath for formatting tables, and ggplot2 for graphs.

Appendix: Consequences of SNP coding 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 monoplots of Items, where "Items" is a generic term for samples such as persons or cultivars. The main text already showed in Table 1 that SNP coding affects the sum of squares (SS) for SNP main effects. Therefore, Item-Centered PCA is not immune because different proportions of main and interaction effects can change which PC is dominated by the SNP main effects, thereby dramatically altering a PCA monoplot of Items. This same verdict of not being immune also applies to Item-Standardized PCA for the same sort of 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 monoplots for Items. Double-Centering pertains to the matrix

$Y_{DC} = \left(I_p - p^{-1}1_p1_p^T\right)Y_C$. 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 $\tilde{Y}_{DC} = \left(I_p - p^{-1}1_p1_p^T\right)PY_C$. The matrices P and $\left(I_p - p^{-1}1_p1_p^T\right)$ do not commute; that is, $\left(I_p - p^{-1}1_p1_p^T\right)P \neq P\left(I_p - p^{-1}1_p1_p^T\right)$ so $\tilde{Y}_{DC} = \left(I_p - p^{-1}1_p1_p^T\right)PY_C \neq P\left(I_p - p^{-1}1_p1_p^T\right)Y_C = PY_{DC}$. 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 SS for Item main effects is small relative to that 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, but not exactly 1. 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 monoplots of Items to be virtually immune to changes in SNP coding.

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: 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.

Supporting information

S1 OatMixed. The oat dataset with SNP coding mixed as received from Kathy Esvelt Klos, except that the original coding of 1 and 2 was transposed to 0 and 1. It has 635 oat lines and 1341 SNPs.

S2 OatRare1. The oat dataset with SNP coding rare=1, which required polarity reversal for 772 of the 1341 SNPs.

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

S4 ThreeTables. Three augmented ANOVA tables for SNP-Standardized, Item-Standardized, and Grand-Mean-Centered PCA of SP data on oats, using SNP coding rare=1.

S6 LiteratureSurvey. Literature survey of 125 articles that apply PCA analysis to SNP data.

S5 FiveBiplots. Five biplots for the oat data: SNP-Centered, SNP-Standardized, Item-Centered, Item-Standardized, and Grand-Mean-Centered PCA using SNP coding rare=1. As in the main text, to reduce clutter, all of these biplots use two panels, with oat lines on the left and SNPs on the right. The color scheme is the same as in Fig 4 and elsewhere in the main text, namely spring oats show in green, world diversity oats shown in blue, and winter oats shown in red, with corresponding colors for the SNPs.

894 ## References

895 1. Gabriel K. The biplot graphic display of matrices with application to principal component
896 analysis. Biometrika. 1971;58:453–467.
897 2. Gower J, Lubbe S, le Roux N. Understanding biplots. New York: John Wiley and Sons; 2011.
898 3. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for
899 association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633-2635.
900 4. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format
901 and VCFtools. Bioinformatics. 2011;27:2156-2158.
902 5. Morrison D, Van Iersel L, Kelk S, List M. Distortions and artifacts in principal components analysis
903 of genome data. 2012. Available from:
904 http://phylonetworks.blogspot.com/2012/12/distortions-and-artifacts-in-pca.html.
905 6. Morrison D, Van Iersel L, Kelk S, List M. Continued misuse of PCA in genomics studies. 2016.
906 Available from: http://phylonetworks.blogspot.com/2016/05/continued-misuse-of-pca-in-
907 genomics.html.
908 7. Gauch HG. Multivariate analysis in community ecology. Cambridge, UK: Cambridge University
909 Press; 1982.
910 8. Digby PGN, Kempton RA. Multivariate analysis of ecological communities. New York: Chapman
911 and Hall; 1987.
912 9. Novembre J, Stephens M. Interpreting principal component analyses of spatial population
913 genetic variation. Nat Genet. 2008;40:646-649.
914 10. Reich D, Price AL, Patterson N. Principal component analysis of genetic data. Nature Genet
915 2008;40:491-492.
916 11. Glantz SA, Slinker BK, Slinker BK. Primer of applied regression & analysis of variance New York:
917 McGraw-Hill Education; 2016.
918 12. Esvelt Klos K, Huang YF, Bekele WA, Obert DE, Babiker E, Beattie AD, et al. Population Genomics
919 Related to Adaptation in Elite Oat Germplasm. Plant Genome. 2016;9.
920 13. Gauch HG, Piepho H-P, Annicchiarico P. Statistical analysis of yield trials by AMMI and GGE:
921 Further considerations. Crop Sci. 2008;48:866-889.
922 14. Travis AJ, Norton GJ, Datta S, Sarma R, Dasgupta T, Savio FL, et al. Assessing the genetic diversity
923 of rice originating from Bangladesh, Assam and West Bengal. Rice (N Y). 2015;8:35.
924 15. Han Y, Zhao X, Cao G, Wang Y, Li Y, Liu D, et al. Genetic characteristics of soybean resistance to
925 HG type 0 and HG type 1.2.3.5.7 of the cyst nematode analyzed by genome-wide association
926 mapping. BMC Genomics. 2015;16:598.
927 16. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, et al. Comprehensive
928 genotyping of the USA national maize inbred seed bank. Genome Biol. 2013;14:R55.
929 17. Kolech SA, Halseth D, Perry K, Wolfe D, Douches DS, Coombs J, et al. Genetic Diversity and
930 Relationship of Ethiopian Potato Varieties to Germplasm from North America, Europe and the
931 International Potato Center. Am J Potato Res. 2016;93:609-619.
932 18. Galinsky KJ, Bhatia G, Loh PR, Georgiev S, Mukherjee S, Patterson NJ, et al. Fast Principal-
933 Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. Am J Hum
934 Genet. 2016;98:456-472.

935 19. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising
936 to the challenge of larger and richer datasets. GigaScience. 2015;4:7.
937 20. Prive F, Aschard H, Ziyatdinov A, Blum MGB. Efficient analysis of large-scale genome-wide data
938 with two R packages: bigstatsr and bigsnpr. Bioinformatics. 2018;34(16):2781-2787. Epub
939 2018/04/05.
940 21. Zeng X, Chakraborty R, King JL, LaRue B, Moura-Neto RS, Budowle B. Selection of highly
941 informative SNP markers for population affiliation of major US populations. Int J Legal Med.
942 2016;130:341-352.
943 22. Zeng X, Warshauer DH, King JL, Churchill JD, Chakraborty R, Budowle B. Empirical testing of a 23-
944 AIMs panel of SNPs for ancestry evaluations in four major US populations. Int J Legal Med.
945 2016;130:891-896.
946 23. Malik WA, Piepho H-P. Biplots: Do not stretch them! Crop Sci. 2018;58:1-9.
947 24. Chen J, Zheng H, Bei JX, Sun L, Jia WH, Li T, et al. Genetic structure of the Han Chinese
948 population revealed by genome-wide SNP variation. Am J Hum Genet. 2009;85:775-785.
949 25. Abraham G, Inouye M. Fast principal component analysis of large-scale genome-wide data. PLoS
950 One. 2014;9:e93766.
951 26. Jackson JE. A user's guide to principal components. New York: Wiley-Interscience; 1991.

952