Supplementary Note: Detecting population structure in rare variant data

Inferring ancestry from genetic data is a common problem in both population and medical genetic studies, and many methods exist to address it [1-3]. Principal components analysis (PCA) [2] has been shown to be effective at elucidating geographic structure from genetic data [4] and correcting for confounding due to population stratification in association mapping [5]. These uses of PCA depend critically on its ability to separate genetically disparate subpopulations when analyzing data from commercial genotyping arrays. However, as high-throughput sequence data becomes more common, enabling ancestry inference from this new class of data is becoming increasingly relevant.

As sequence data contains more variants, and many more population-specific variants [6], it may be reasonable to expect that PCA applied to high-throughput sequence data will be substantially more effective than the corresponding analysis on genotype data. However, our results suggest the opposite. Specifically, PCA makes assumptions about marker independence that are violated by the pervasive linkage disequilibrium in sequence data. In addition, assumptions about genetic drift that are reasonable for common SNPs on genotyping arrays are less so when applied to the numerous rare variants in sequence data [7].

Principal Components Analysis (PCA) is generally applied to a genetic relationship matrix (GRM) that is computed as:

$$g_s = \frac{x_s - 2p_s}{\sqrt{2p_s(1-p_s)}}$$

$$G = \sum_{s \in SNPs} g_s g_s^T$$

where x_s is a vector of genotypes for SNP s and p_s is the minor allele frequency of SNP s. We propose modifications to this GRM to deal with two challenges that are present in sequence data but absent from genotype data: pervasive linkage disequilibrium, and rare variants. Specifically, we recommend that LD pruning be applied to sequence data and singleton variants be removed. While we evaluated more sophisticated approaches to handling these issues, they did not improve our results beyond these simpler approaches. Importantly, we recommend against a commonly used strategy of removing all low frequency of rare variants as these variants contain significant information for detecting population structure.

Linkage Disequilibrium

It is well known that application of PCA to regions of the genome containing long-range LD blocks can confound PCA's ability to separate disparate populations [2,8]. As a result, these LD blocks are often simply excluded from analysis. However, in sequence data, many regions of the genome outside of previously identified long-range LD blocks contain sufficient LD to bias results. As a result, we examine three methods to deal with LD: (1) LD Pruning (2) LD Shrinkage[8] and (3) LD Regression[2,9].

LD Pruning is a commonly applied approach to removing correlated SNPs from a dataset. To produce a data set pruned for LD above a threshold T, one SNP of any pair of SNPs in LD ($r^2 > T$) is removed from the data.

LD Shrinkage is a more sophisticated method of correcting for LD proposed by (Zou et al. 2012). In LD shrinkage, each SNP s is weighted by its LD to surrounding SNPs before inclusion in the genetic relationship matrix:

$$g_s = \frac{x_s - 2p_s}{\sqrt{2p_s(1-p_s)}}$$
$$w_s = \frac{1}{1 + \sum_{t \in window(s)} r_{s,t}^2}$$
$$G = \sum_{s \in SNPs} g_s g_s^T$$

We note that $t \in window(s)$ refers to SNPs t that are within some region of the genome surrounding SNP *s*. Intuitively, this is a heuristic to correct for the over representation in the GRM of some SNPs that are redundant with respect to nearby SNPs.

LD Regression was originally proposed in (Patterson et al. 2006) and essentially involves the inclusion of only "residualized" SNPs in the GRM:

$$g_s = \frac{x_s - 2p_s}{\sqrt{2p_s(1-p_s)}}$$
$$g_s \sim \sum_{t \in window(s)} g_t + \varepsilon_s$$
$$G = \sum_{s \in SNPs} \varepsilon_s \varepsilon_s^T$$

Rare Variants

In considering how to optimally include rare variants in the genome, we examined three strategies. The first strategy was to include all rare variants as described in the computations above without any modifications. The second strategy was to exclude all variants below a threshold, which is a standard strategy used in several recent papers. We compared these simple strategies to a strategy based on reweighting rare variants to optimize the separation between populations.

We considered a particular scenario to optimize. Specifically, we imagine that two populations that split from one another t generations ago are equally represented in our GRM. We would like to optimize the proportion of variance in our GRM that is explained by the true population labels. That is, our figure of merit is:

$$\frac{\frac{1}{n(n-1)} \sum_i \sum_{j \in pop(i)} g_{i,j} - \frac{1}{n^2} \sum_i \sum_{j \in pop(i)} g_{i,j}}{\sqrt{\mathrm{Var}(g_{i,j})}}$$

where $pop(i)$ refers to the subpopulation from which individual i came.

Now, considering the population split, our data contains two classes of variants: those variants that are result of mutations predating the population split (pre-split SNPs), and those variants arising after the population split (post-split SNPs). Now, for pre-split SNPs

we invoke a normal approximation to genetic drift. That is, the difference between allele frequencies p_1, p_2 (for populations 1 and 2, respectively) is:

$$(p_1 - p_2) \sim N\left(0, 2F_{ST}p(1-p)\right)$$

where p is the allele frequency in the ancestral population prior to the split and F_{ST} quantifies the genetic drift that has occurred since the split. We note that this approximation is reasonable for common SNPs and for small values F_{ST}. If we assume that our data contains only pre-split SNPs then our figure of merit is optimized by the standard computation of the GRM given above. However, if we assume that our data also contains rare, post-split SNPs then our optimal GRM is different. These variants have the property that

$$|p_1 - p_2| = 2\hat{p}$$

where $\hat{p}$ is the allele frequency estimated from the sample. This is because post-split SNPs have a population allele frequency of exactly 0 in one of the two populations studied (ignoring recurrent mutation). In this context, we continue to treat pre-split SNPs identically:

$$g_i^s = \frac{x_i - 2p_s}{\sqrt{p_s(1-p_s)}}, \text{for pre} - \text{split SNP } s$$

but

$$g_i^s = (x_i - 2p_s)\sqrt{\frac{F_{ST}^2 + 2F_{ST} + 2}{F_{ST}(1-2p_s)}}, \text{for post} - \text{split SNP } s$$

However, this modification requires knowledge of the F_{ST} between studied subpopulations and, more dauntingly, which SNPs are post-split. We believe it is reasonable to iterate over several values of F_{ST} (and find that in real data results are relatively robust to choice of F_{ST}). In order to deal with uncertainty over the set of post-split SNPs, we propose that a SNP be considered post-split if

$$\frac{1}{\sqrt{p_s(1-p_s)}} > \sqrt{\frac{F_{ST}^2 + 2F_{ST} + 2}{F_{ST}(1-2p_s)}}$$

We examine the effect of both of these modifications on the effectiveness of PCA to separate genetically disparate subpopulations.

Analysis of Northern vs. Southern Europe in POPRES Targeted Sequencing Data

We analyzed 531 individuals from the UK referred to as Northern European and 146 Italian, 134 Portuguese, 100 Spaniards, and 7 Swiss Italian individuals collectively referred to as Southern European[10]. We excluded 25.9 kb of sequence data from genes on the X chromosome, focusing solely on the autosomes. In total, 8,469 SNPs were polymorphic in either of the Northern or Southern European Samples. These variants were overwhelmingly rare, with 81.5% of variants having a MAF < 1% in the combined sample.

We tested various methods to correct for LD and better handle rare variants (see Methods). The results are summarized in Supplementary Table 11. These results indicate that handling of both rare variants and LD is critical to maximizing the performance of PCA on this class of data. Applying standard PCA, the top 5 PCs explained only 2.3% of the variance (r^2=0.023) of the true population labels. This was improved substantially by

removing or reweighting rare variants with (r^2=0.287, 0.341, 0.352) for removing variants with MAF < 0.02, removing singletons and reweighting, respectively. This indicates that rare variants, particularly singletons, may be problematic when analyzed using PCA. However, the difference between removing variants with MAF < 0.02 and reweighting (r^2=0.287 vs 0.352) suggests that these variants do contain useful information for ancestry inference and should not be universally excluded.

Additionally, application of a method to correct for LD significantly improved performance of PCA when performed in conjunction with singleton exclusion or rare variant reweighting. With rare variant reweighting, LD shrinkage [8] (r^2=0.563) performing slightly better than LD regression (r^2=0.528) [2] and LD pruning (r^2=0.534). While LD Pruning performed well, this may be due to the fact that LD is broken up because the dataset contains sequence data from separated chunks of genome.

Recommendations

In data sets that do not include pervasive LD or large numbers of rare variants (i.e. genotyping data), standard techniques are likely to be successful in detecting population structure. However, in data sets that have pervasive LD and large numbers of rare variants, we recommend that LD pruning and singleton removal be applied. While more sophisticated methods for dealing with these issues were assessed, we did not observe significant improvements above and beyond these simpler approaches. Importantly, we do not recommend that all low frequency and rare variants (MAF < 0.02) be removed as these variants do significantly improve detection of population structure.

References

1. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).
2. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2,** e190 (2006).
3. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of Population Structure using Dense Haplotype Data. *PLoS Genet* **8,** e1002453 (2012).
4. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (2008).
5. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38,** 904–909 (2006).
6. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
7. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nature genetics* **44,** 243–246 (2012).
8. Zou, F., Lee, S., Knowles, M. R. & Wright, F. A. Quantification of population structure using correlated SNPs by shrinkage principal components. *Hum. Hered.* **70,** 9–22 (2010).
9. Gusev, A. *et al.* Quantifying Missing Heritability at Known GWAS Loci. *PLoS Genet* **9,** e1003993 (2013).
10. Nelson, M. R. *et al.* An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *Science* **337,** 100–104 (2012).

11. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature genetics* **43,** 1193–1201 (2011).

Supplementary Table 11. Evaluation of LD and rare variant strategies for running PCA in POPRES targeted sequencing data. We evaluated four methods for dealing with LD, and four methods for dealing with rare variants. We report the total variance explained by the top PCs in distinguishing Northern and Southern Europeans in POPRES targeted sequencing data.

		LD Strategy			
		Standard PCA	LD Pruning	LD Shrinkage	LD Regression
Rare Variant Strategy	Include all variants	0.023	0.012	0.007	0.006
	Exclude MAF < 0.02	0.287	0.441	0.442	0.463
	Exclude Singletons	0.341	0.541	0.567	0.504
	Reweight F_{ST} = 0.01	0.352	0.534	0.563	0.528