PLOS ONE

RESEARCH ARTICLE

Single Nucleotide Polymorphism Clustering in Systemic Autoimmune Diseases

Thomas Charlon[1,2], Manuel Martínez-Bueno[3], Lara Bossini-Castillo[4], F. David Carmona[4], Alessandro Di Cara[1], Jérôme Wojcik[1]*, Sviatoslav Voloshynovskiy[2], Javier Martín[4☯], Marta E. Alarcón-Riquelme[3☯]

1 Quartz Bio, Plan-les-Ouates, Geneva, Switzerland, **2** Stochastic Information Processing, University of Geneva, Geneva, Geneva, Switzerland, **3** Center for Genomics and Oncological Research: Pfizer/University of Granada/Andalusian Government, Granada, Granada, Spain, **4** Institute of Parasitology and Biomedicine López Neyra, Spanish National Research Council, Armilla, Granada, Spain

☯ These authors contributed equally to this work.
* jerome.wojcik@quartz.bio

CrossMark
click for updates

OPEN ACCESS

Citation: Charlon T, Martínez-Bueno M, Bossini-Castillo L, Carmona FD, Di Cara A, Wojcik J, et al. (2016) Single Nucleotide Polymorphism Clustering in Systemic Autoimmune Diseases. PLoS ONE 11(8): e0160270. doi:10.1371/journal.pone.0160270

Editor: Anna Carla Goldberg, Hospital Israelita Albert Einstein, BRAZIL

Received: January 13, 2016

Accepted: July 15, 2016

Published: August 4, 2016

Copyright: © 2016 Charlon et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data are available from the Genyo and CSIC institutions for researchers who meet the criteria for access to confidential data. Access to data can be requested to Dr. Marta Alarcòn- Riquelme (marta.alarcon@genyo.es).

Funding: Quartz Bio S.A. provided support in the form of salaries for TC, ADC, and JW, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. This work has received support from the EU/EFPIA/ Innovative Medicines Initiative Joint Undertaking PRECISESADS grant no. 115565.

Abstract

Systemic Autoimmune Diseases, a group of chronic inflammatory conditions, have variable symptoms and difficult diagnosis. In order to reclassify them based on genetic markers rather than clinical criteria, we performed clustering of Single Nucleotide Polymorphisms. However naive approaches tend to group patients primarily by their geographic origin. To reduce this "ancestry signal", we developed SNPClust, a method to select large sources of ancestry-independent genetic variations from all variations detected by Principal Component Analysis. Applied to a Systemic Lupus Erythematosus case control dataset, SNPClust successfully reduced the ancestry signal. Results were compared with association studies between the cases and controls without or with reference population stratification correction methods. SNPClust amplified the disease discriminating signal and the ratio of significant associations outside the *HLA* locus was greater compared to population stratification correction methods. SNPClust will enable the use of ancestry-independent genetic information in the reclassification of Systemic Autoimmune Diseases. SNPClust is available as an R package and demonstrated on the public Human Genome Diversity Project dataset at https://github.com/ThomasChln/snpclust.

Introduction

The PRECISESADS project aims at reclassifying Systemic Autoimmune Diseases (SADs), a group of chronic inflammatory conditions characterized by the presence of unspecific autoantibodies in the serum and serious clinical consequences, based on genetic and molecular biomarkers rather than clinical criteria. SADs affect 1% of the global population [1] and have limited treatment options and difficult diagnosis. The diseases studied in PRECISESADS are Systemic Lupus Erythematosus (SLE), Systemic Sclerosis, Rheumatoid Arthritis, Sjögren's Syndrome, Primary Antiphospholipid Antibody Syndrome, and undifferentiated cases.

Competing Interests: TC, ADC, and JW are employees of Quartz Bio S.A., Switzerland. The authors declare no competing interests related to this commercial affiliation. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

Several technological platforms are used to generate biomarker data from patient samples, to obtain as much information as possible about the genetic and molecular mechanisms involved. The technologies include Single Nucleotide Polymorphisms (SNPs) microarrays, which measures hundreds of thousands of common genetic variations in a population, gene expression and protein microarrays, mass spectrometry for metabolic profiling, or flow cytometry, as examples. The methodological approach is to first analyze data from each technological platform individually and then to merge relevant features from each platform to build the final classifier. Here we present the results of the preparatory work for SNP clustering analysis performed on a test dataset.

Familial and twin studies have estimated a 50% genetic component in SADs [2, 3] and Genome-Wide Association Studies (GWAS) have found several loci associated with SADs [4]. However, genome-wide clustering of SNPs is known to primarily group patients by ancestry prior to disease relevant features [5, 6]. In order to emphasize the disease relevant signal, we developed SNPClust, a clustering method to overcome this "ancestry bias" by selecting and summarizing SNPs contributing strongly to localized sources of genetic variation as detected by Principal Component Analysis (PCA) [7].

SNPClust first applied PCA to project patients on the largest sources of variance by linear combinations of SNPs. Then for each principal component, the SNPs that had significantly high contributions were selected. Many correlated SNPs were selected from specific loci due to linkage disequilibrium between SNPs, which form haplotypes, and therefore still produced an ancestry signal. To address this, for SNPs selected from the same principal component, SNPClust summarized physically close SNPs in linkage disequilibrium by one variable inferring a haplotype, and reduced the ancestry signal while conserving the other underlying genetic signals.

The test dataset contained SNP microarray data from 4,212 European SLE patients and 1,221 European controls. After quality control, Minor Allele Frequency (MAF) filtering, and tag SNP selection [8], PCA was performed on 379,190 SNPs from 5,433 patients. For each of the 100 first principal components, strong contributing SNPs were selected and SNP-dense regions were summarized by haplotypes. A total of 261 SNPs were selected and 331 haplotypes inferred.

On the SNPClust selected dataset, the clustering signal due to ancestry was significantly reduced. The performance of SNPClust was compared to GWAS standard approaches and reference population substructure correction methods [9, 10]. SNPClust was shown to enrich the selection of ancestry-independent sources of genetic variation associated with the phenotype, and hence propose more robust candidate biomarkers.

Results

Ancestry clusters

As expected, PCA applied on the input dataset (without any prior feature selection or data transformation) grouped patients by the country of origin of the samples, discriminating Northern from Southern Europeans in the first principal component and Eastern from Western Europeans in the second (Fig 1).

This ancestry signal can also be seen in most of the 10 first principal components. The most important non-ancestry-based source of genetic variation in the PCA appeared on principal component 3 (see below), thus confirming that the ancestry signal is much stronger than clinically relevant signals in clustering approaches.

Selection of strong contributors

The analysis of the contributions of SNP markers to the PCA principal component axes revealed that the first twenty principal components were driven by large localized SNP groups.

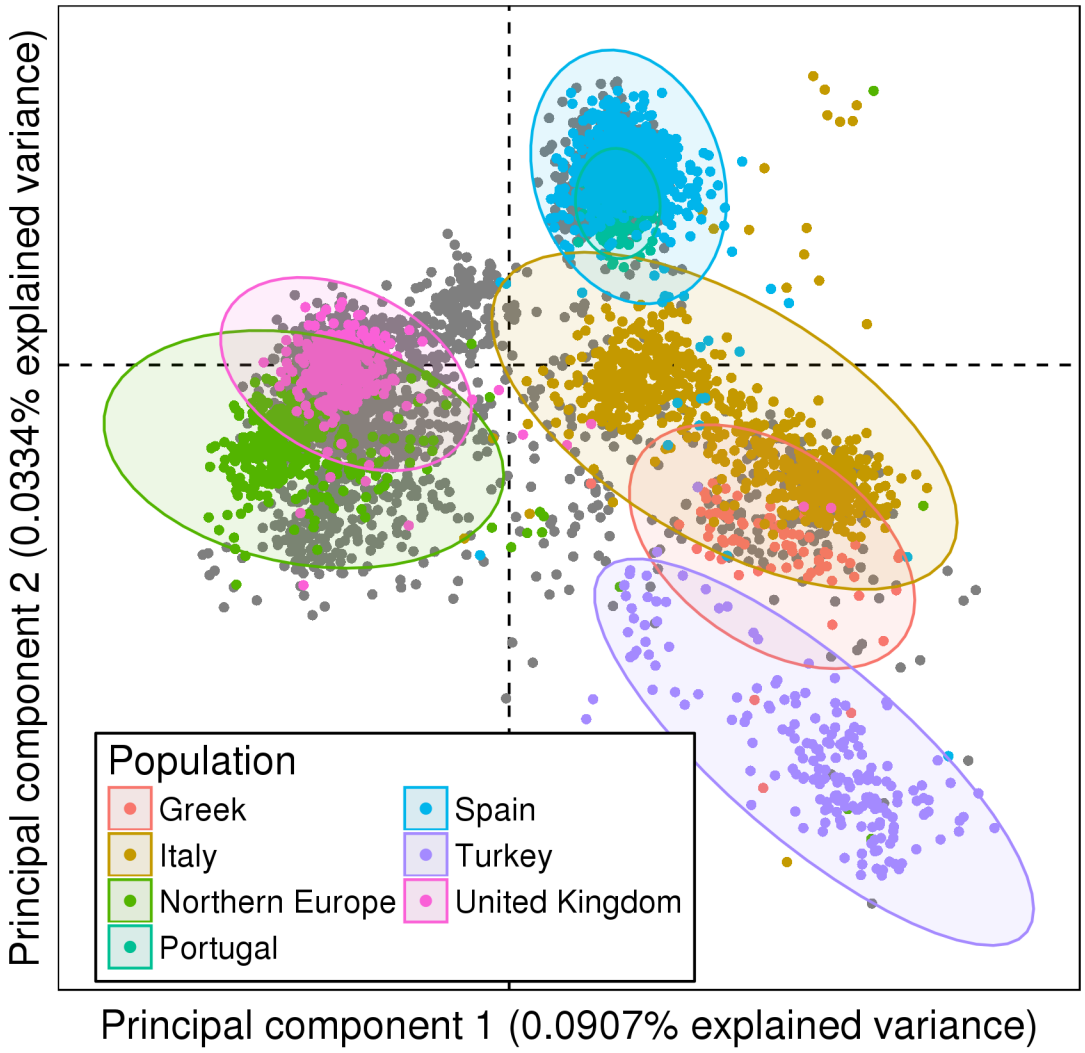

Fig 1. Initial grouping of genetic data. Two first principal components of the PCA on 379,190 SNPs from 5,433 European SLE patients and controls with 95% confidence ellipses. Northern and Southern Europeans were discriminated in the first principal component. Eastern and Western Europeans were discriminated in the second. 2,733 individuals (50%) did not have geographic information and were colored in gray.

doi:10.1371/journal.pone.0160270.g001

The chromosome 6 *HLA* locus was the strongest and largest contributor in all of the first 8 principal components except principal component 3. Principal component 3 was driven by the chromosome 8 locus *8p23*, one of the largest inversion polymorphism encompassing 4,500,000 base pairs and including the SLE associated gene *BLK* [11] (Fig 2). As *8p23* was a main contributor in some of the first principal components, we confirmed that ancestry-independent signals can be extracted by SNPClust.

The 100 first principal components, on which the Gaussian mixture models based selection was applied, explained 3.5% of the total variance. Large localized SNP groups were selected along with other strong contributors. In total, 10,422 SNPs were selected, including 4,090 SNPs from the first 8 principal components (Fig 2).

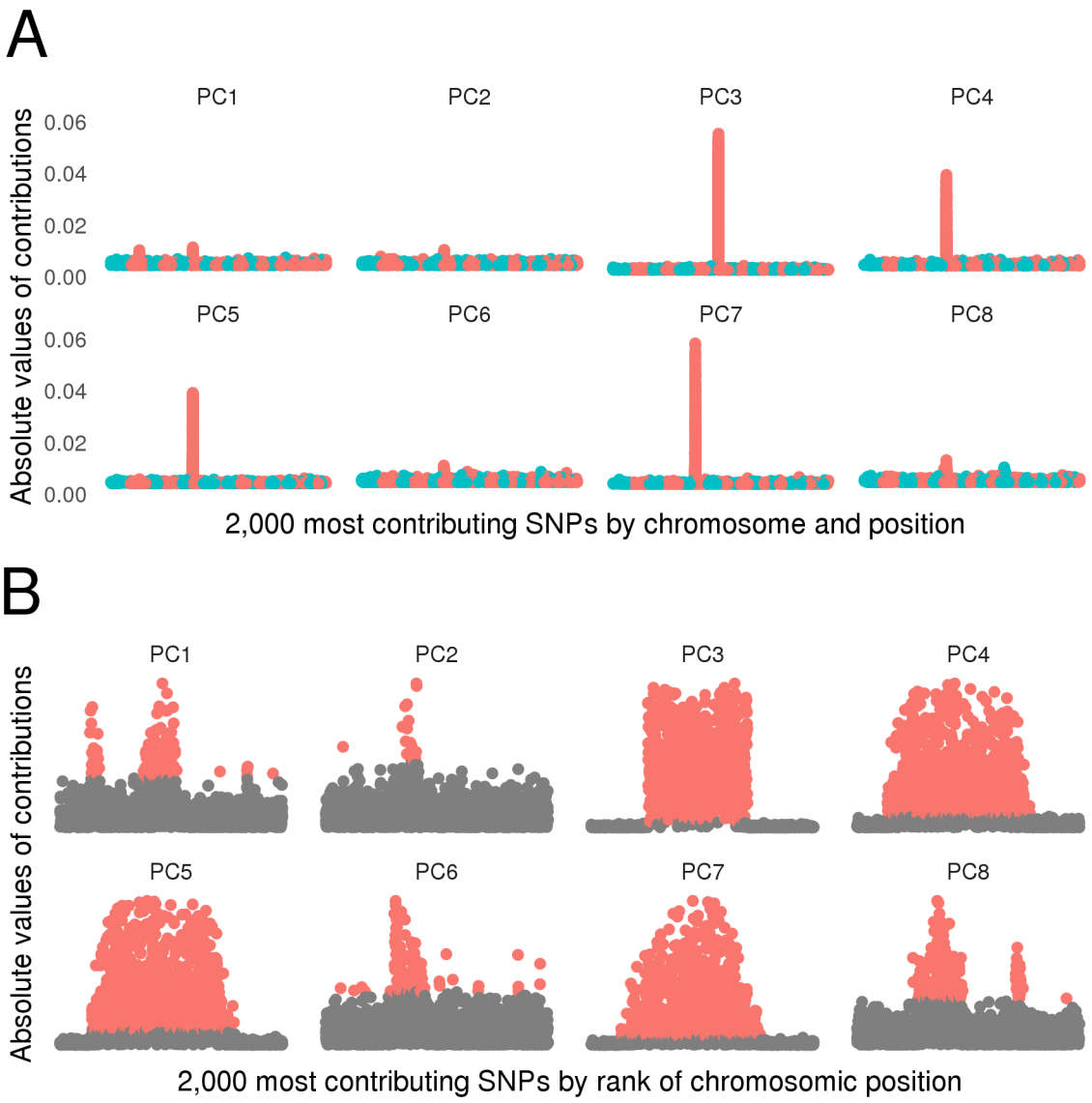

Fig 2. Selection of strong SNP contributors. (a) The 2,000 most contributing SNPs to each of the first 8 principal components are displayed by chromosomal position and colored by chromosomes. Principal components were driven by large localized SNP groups and the chromosome 6 locus *HLA* was the strongest and largest contributor in all of the first 8 principal components, except principal component 5. **(b)** Selection of SNPs by the Gaussian mixtures based method. Selected SNPs are colored in red. SNPs are displayed on the x-axis by rank of chromosomic position, *i.e.* SNPs are regularly spaced and ordered by chromosome and position.

doi:10.1371/journal.pone.0160270.g002

Haplotype summarization

Then, the 10,422 selected SNPs were reduced to 261 SNPs and 331 haplotypes by the haplotype estimation process described in the methods. For example, on principal component 3, the 875 SNPs from chromosome 8 were summarized by 2 distinct haplotypes h1 and h2. Each subject was assigned a h1/h1, h1/h2, or h2/h2 haplotype combination for this region (Fig 3).

Ancestry signal reduction

We applied PCA on the selected SNPs and haplotypes. The first principal components did not cluster patients neither by phenotype nor by centers (Fig 4).

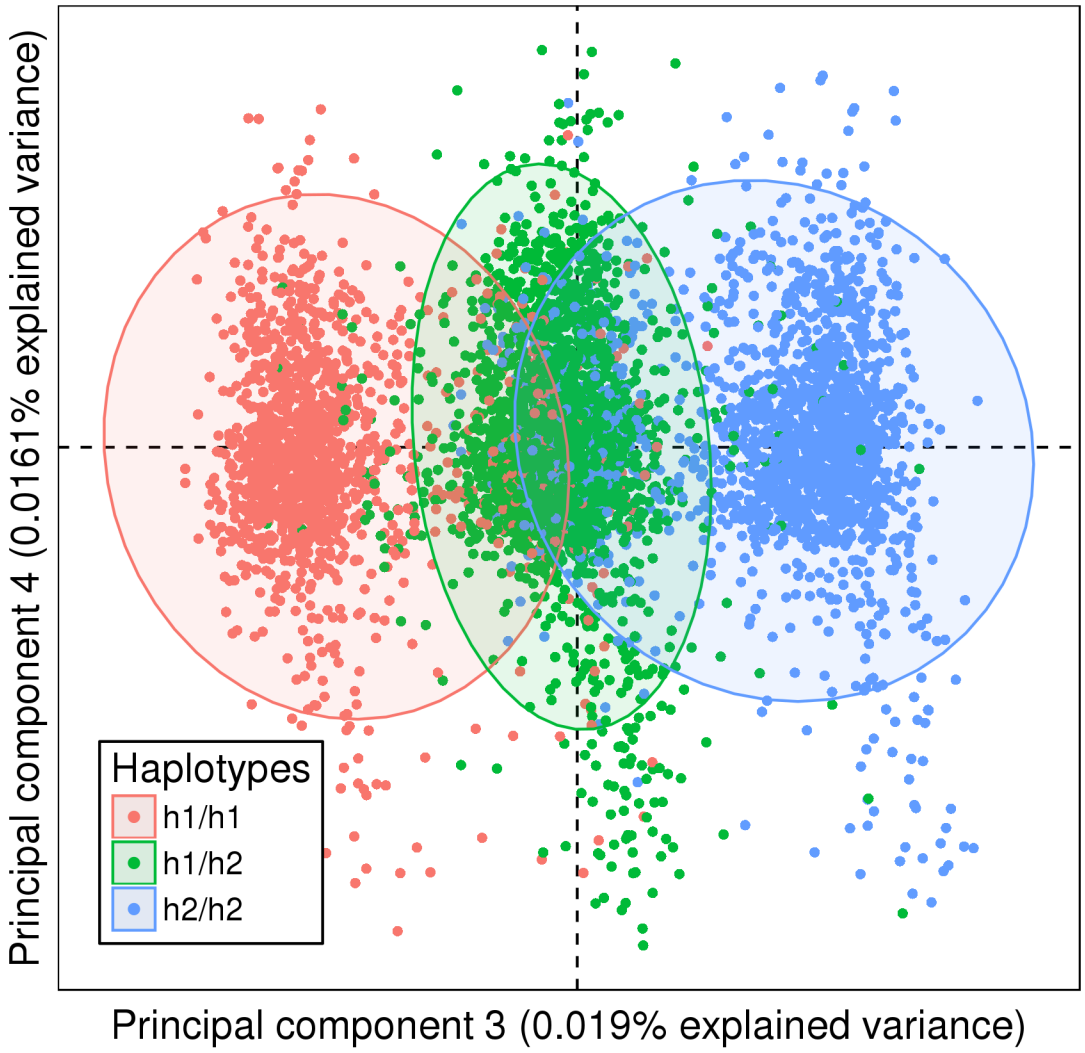

Fig 3. Haplotype summarization of the *8p23* locus. The haplotypes estimated from the 875 selected SNPs from chromosome 8 were best fitted by two groups. The resulting three groups, plotted with 95% confidence ellipses, accurately represented the three clusters in principal component 3 and showed that haplotypes preserved information carried by SNPs.

doi:10.1371/journal.pone.0160270.g003

The method therefore reduced the ancestry signal and produced a set of features that could be further investigated for disease relevant signals.

Benchmark

GWAS have found several loci associated with SADs. However such approaches are impaired by population substructures that can generate false-positives or domestic association signals [4]. In order to evaluate the information conserved by SNPClust and compare it to existing population stratification correction methods, a GWAS was performed on the input dataset, without or with subpopulation structure correction (see methods), and compared with the results of a GWAS performed on the SNPClust selected dataset.

Compared with the GWAS performed directly on the input dataset, the GWAS performed on the SNPClust-processed dataset selected much less SNPs with a nominal p-value < 0.05:

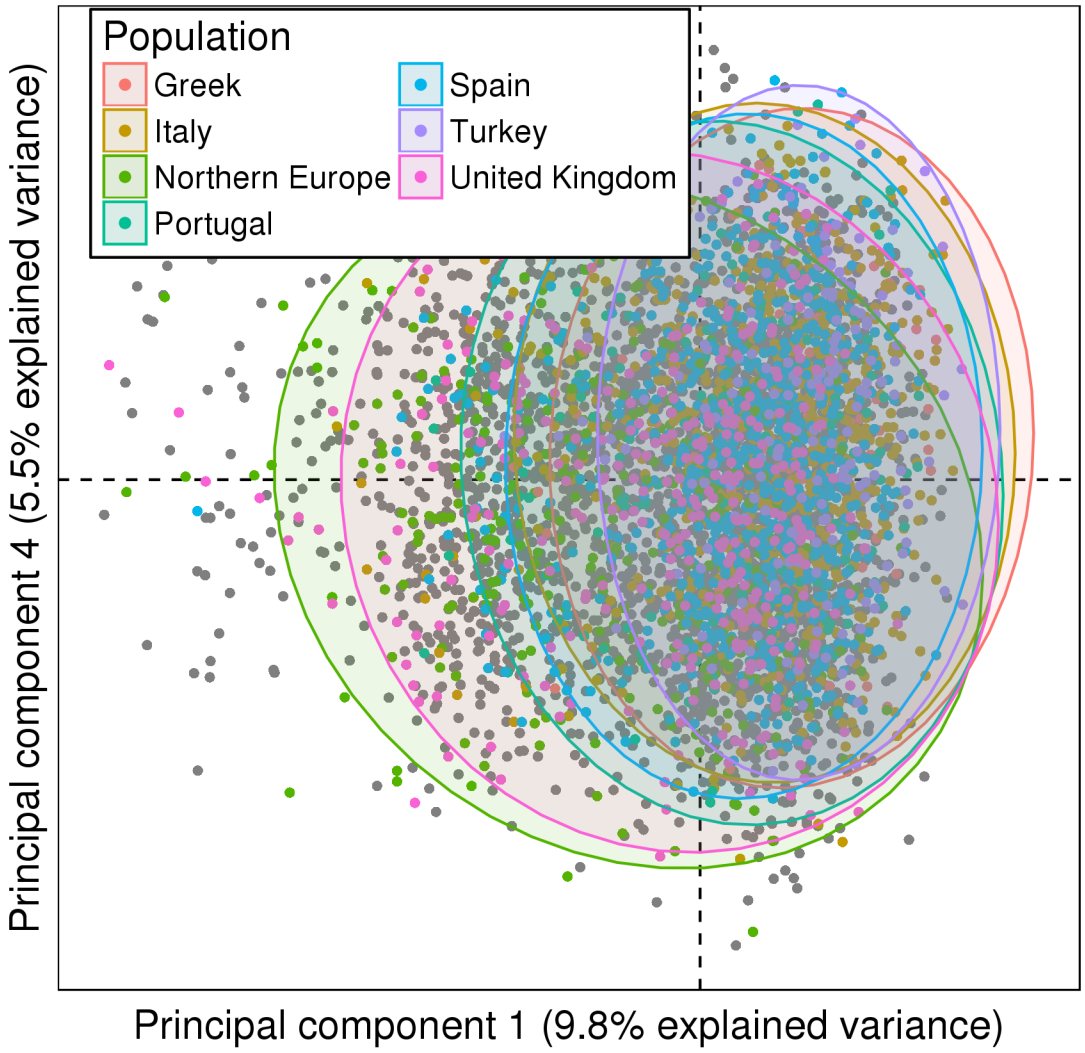

Fig 4. PCA after SNPClust application. First principal components of the PCA on the 261 SNPs and 331 haplotypes from 5,433 patients, with 95% confidence ellipses. The principal components did not discriminate SLE patients from controls.

doi:10.1371/journal.pone.0160270.g004

97,066 *vs.* 222. In addition, the initial selection of SNPs by SNPClust reduced the impact of the multiple testing correction: 59% of SNPs on the input dataset with $p < 0.05$ had a False Discovery Rate (FDR) [12] > 5%, compared to only 30% after SNPClust (Table 1).

Compared with GWAS with genomic control [9], SNPClust produced 42% less significant associations after multiple testing correction, but 3 times more outside the *HLA* region.

Table 1. Performances of feature selection methods. Associations of the input dataset without or with population stratification correction by genomic control and Eigenstrat (with 5 and 10 principal components considered) compared with SNPClust.

Feature selection method	Number of p-values < 5%	Number of FDR q-values < 5%	Number of FDR q-values in *HLA* < 5%
(none)	97,066	39,555	951
Genomic control	18,267	271	232
Eigenstrat 5	22,605	117	64
Eigenstrat 10	15,269	28	9
SNPClust	222	156	38

doi:10.1371/journal.pone.0160270.t001

Therefore SNPClust has less statistical power than genomic control but is much more effective in finding associations from several loci. Eigenstrat [10] was performed with 5 and 10 principal components. In both cases, SNPClust had more statistical power, with less hits in *HLA* (24% *vs.* 55% and 32%).

Increasing statistical power can increase false positive rate. However SLE has still a large unexplained heritability and increasing the power is a possible step to reduce it by discovering novel markers. Additionally, the FDR multiple testing correction method can also be more stringent to reduce the power and the false positive rate.

The *HLA* locus exhibited a pitfall of GWAS with multiple testing correction: even after TagSNP selection, most associated SNPs were from few loci, due to haplotypes. This had the effect of excluding other loci when multiple testing correction was applied. SNPClust overcame this by summarizing the SNPs from one locus in one haplotype. Associated genes found by SNPClust and Eigenstrat with 5 components were compared. SNPClust had 67 unique genes associated, Eigenstrat had 37, and with an intersection of 1. The associated genes were also compared to previously known genes associated with SADs. The intersection had 1 known gene (*NOTCH4*). Excluding this one, SNPClust had 4 known genes (*MICA*, *MSH*, *PSORS1C1*, *RAD51B*) and Eigenstrat 4 (*ATG5*, *BANK1*, *STAT4*, *TNXB*).

Discussion

The ancestry information is contained in many SNPs across the genome, and may therefore be present in clinically relevant SNPs, in particular in auto-immune diseases where the *HLA* locus is involved. Therefore removing simply the main known ancestry-informative markers may lead to the removal of clinically relevant SNPs while preserving many SNPs carrying small bits of ancestry information. Approaches considering the first principal components to adjust associations [10] can also result in loss of clinically relevant information because not all the first principal components are associated with ancestry.

SNPClust overcomes these limitations in two steps. First the major contributors to a large number of the first principal components are considered, therefore selecting the markers explaining most of the variance in the dataset. This has the property of preserving the largest sources of genetic variation. Then, the SNPs that could be considered as haplotypes due to their correlation and spatial proximity are summarized. This reduces the relative importance of ancestry information present in many SNPs while preserving the information conveyed by the haplotypes. At the same time, it also reduces the multiple-testing problem.

In the dataset of 379,190 SNPs from SLE patients and controls, the first principal components were associated with ancestry. After application of SNPClust 592 markers remained and the first principal components were not associated to the centers. Therefore the ancestry information and the number of markers were strongly reduced.

This method is also interesting for GWAS, due to the enrichment of ancestry-independent markers tested and the reduced multiple testing problem. When compared to genomic control, 42% less SNPs were associated with the diseases but 10 times more outside the *HLA* locus. Compared to Eigenstrat, SNPClust had more statistical power and a higher ratio of associations outside *HLA*. Therefore SNPClust is useful to find several associated loci without being overwhelmed by strong signals from one single locus such as *HLA*.

SNPClust will be applied to the SNP array data generated in the PRECISESADS project and will possibly enable the use of ancestry-independent genetic information in the reclassification of SADs. It is available as an R package and demonstrated on the public Human Genome Diversity Project dataset [6] at https://github.com/ThomasChln/snpclust.

Materials and Methods

Input dataset

The dataset contained 4,212 European SLE patients and 1,221 European healthy controls. It was previously published [13] and approved by ethics committees. No samples were used and records were de-identified. The files were binary PLINK files [14]. They were converted in the Genomic Data Structure format [15].

Analysis set

Observations with missing value rates above 3% and SNPs with missing value rates above 1% were excluded. An additive genetic model is then used (AA = 0, AB = 1, BB = 2) and SNPs with MAF below 5% were excluded to remove rare variants, which are more prone to genotyping errors. In addition, in order to decrease the required computation time and memory usage, redundant SNPs were removed by applying TagSNP ($r^2 > 0.8$, window of 500,000 base pairs). The missing values were imputed by random sampling of each SNP. Then each SNP was centered and scaled to unit variance.

A total of 5,433 patients and 379,190 SNPs were selected for analysis. This dataset defines our analysis set.

SNPClust

Selection of strong contributors to principal components. PCA was applied on the analysis set. PCA is a dimensionality reduction method, which projects SNPs by linear combination to maximize the variance on successive axes, *i.e.* principal components, while constraining the axes to be orthogonal. SNPs with large absolute projection values, *i.e.* loadings or contributions, to the 100 first principal components were selected.

For each principal component, a Gaussian mixture model [16] with 2 mixture components was fitted to the absolute values of SNP contributions. Only the 3,000 highest absolute contributions were considered for computational performance. In Gaussian mixtures, SNPs have a probability of being assigned to each Gaussian model, from which can be derived a classification uncertainty. Only the strong contributors have null uncertainty, therefore SNPs with a null classification uncertainty were selected. If SNPs were selected from more than 8 chromosomes, the model was fitted with an additional mixture until the condition was satisfied. If the condition was not satisfied after 4 iterations, *i.e.* with 5 mixtures, no SNPs were selected from that principal component (Algorithm 1).

Algorithm 1 Selection of strong contributors

Input: PCA rotation matrix *i.e.* SNPs coefficients to principal components
for *coefficients* **in** $coefficients_{PC1}, \ldots, coefficients_{PC100}$ **do**
  coefficients ← 3,000 highest absolute values of *coefficients*
  selection ← ∅
  for $n_{mixtures}$ **in** 2, 3, 4, 5 **do**
    GMM ← Gaussian mixture model of $n_{mixtures}$ components on *coefficients*
    selection ← *coefficients* with null uncertainty classification in *GMM*
    if Number of chromosomes in *selection* < = 8 **then**
      exit for loop
    end if
  end for
  store *selection* **in output**
end for
Output: $selection_{PC1}, \ldots, selection_{PC100}$

Haplotype summarization. In a second step, in order to summarize large loci, for each principal component, selected SNPs on a same chromosome and closer than 1,000,000 base pairs away were summarized by haplotypes, pairs of binary values. Haplotypes were inferred with the SHAPEIT software [17] which uses hidden Markov models and has linear complexity with the number of SNPs. Haplotypes were then grouped in 2 classes by Gaussian mixture models. Correlated haplotypes were removed when they were on the same chromosome and the squared correlation was above 0.8 (Algorithm 2).

Algorithm 2 Summarization of physically close SNPs

```
Input: Selected SNPs for the 100 first principal components
haplotypes ← ∅
for SNPs in selection_PC1, ..., selection_PC100 do
  for locus in groups of SNPs closer than 1,000,000 base pairs in SNPs do
    haplotype ← Haplotype estimation of locus
    SNPs ← Exclude locus from SNPs
    store haplotype in haplotypes
  end for
end for
for haplotypes in groups of haplotypes from same chromosome do
  haplotypes ← Squared correlation threshold of 0.8 on haplotypes
end for
Output: Union of haplotypes and SNPs
```

Benchmark

To evaluate the SNPClust algorithm, we performed a GWAS on the SNPClust selected dataset and on the input dataset with and without population stratification correction. First, a generalized linear model with a binomial error distribution was fitted to each SNP and haplotype to predict the disease of patients. Type 2 Analysis of Variance was then applied to obtain p-values.

Multiple testing correction was performed with FDR, and results were compared with the outcome of our method. Then, two population stratification correction methods were tested, genomic control and Eigenstrat.

Author Contributions

Conceptualization: JW SV.

Data curation: LBC MEAR.

Formal analysis: TC.

Funding acquisition: MEAR JW.

Investigation: MEAR JW ADC.

Methodology: TC JW.

Project administration: JW.

Resources: MMB LBC FDC JM MEAR.

Software: TC.

Supervision: JW SV.

Validation: TC JW ADC.

Visualization:TC.

Writing - original draft: TC.

Writing - review & editing: JW ADC MEAR.

References

1. Cooper GS, Bynum ML, Somers EC. Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases. Journal of autoimmunity. 2009; 33(3):197–207. doi: 10.1016/j.jaut.2009.09.008 PMID: 19819109
2. Cooper GS, Miller FW, Pandey JP. The role of genetic factors in autoimmune disease: implications for environmental research. Environmental Health Perspectives. 1999; 107(Suppl 5):693. doi: 10.2307/3434329 PMID: 10502533
3. Bax M, van Heemst J, Huizinga TW, Toes RE. Genetics of rheumatoid arthritis: what have we learned? Immunogenetics. 2011; 63(8):459–466. doi: 10.1007/s00251-011-0528-6 PMID: 21556860
4. Delgado-Vega A, Sánchez E, Löfgren S, Castillejo-López C, Alarcón-Riquelme ME. Recent findings on genetics of systemic autoimmune diseases. Current opinion in immunology. 2010; 22(6):698–705. doi: 10.1016/j.coi.2010.09.002 PMID: 20933377
5. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008; 456(7218):98–101. doi: 10.1038/nature07331 PMID: 18758442
6. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. science. 2008; 319(5866):1100–1104. doi: 10.1126/science.1153717 PMID: 18292342
7. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genet. 2006; 2(12): e190. doi: 10.1371/journal.pgen.0020190 PMID: 17194218
8. Stram DO. Tag SNP selection for association studies. Genetic epidemiology. 2004; 27(4):365–374. doi: 10.1002/gepi.20028 PMID: 15372618
9. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55(4):997–1004. PMID: 11315092
10. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006; 38(8):904–909. doi: 10.1038/ng1847 PMID: 16862161
11. Namjou B, Ni Y, Harley IT, Chepelev I, Cobb B, Kottyan LC, et al. The Effect of Inversion at 8p23 on BLK Association with Lupus in Caucasian Population. PloS one. 2014; 9(12):e115614. doi: 10.1371/journal.pone.0115614 PMID: 25545785
12. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995; p. 289–300.
13. Bentham J, Morris DL, Graham DSC, Pinder CL, Tombleson P, Behrens TW, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nature genetics. 2015;.
14. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007; 81(3):559–575. doi: 10.1086/519795 PMID: 17701901
15. Zheng X, Levine D, Shen J, Gogarten S, Laurie C, Weir B. A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. Bioinformatics. 2012; 28(24):3326–3328. doi: 10.1093/bioinformatics/bts606 PMID: 23060615
16. Fraley C, Raftery AE. Model-based Clustering, Discriminant Analysis and Density Estimation. Journal of the American Statistical Association. 2002; 97:611–631. doi: 10.1198/016214502760047131
17. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nature methods. 2013; 10(1):5–6. doi: 10.1038/nmeth.2307 PMID: 23269371