OPENING THE DOOR TO THE LARGE SCALE USE OF CLINICAL LAB MEASURES FOR ASSOCIATION TESTING: EXPLORING DIFFERENT METHODS FOR DEFINING PHENOTYPES

CHRISTOPHER R. BAUER[1], DANIEL LAVAGE[1], JOHN SNYDER[1], JOSEPH LEADER[1], J. MATTHEW MAHONEY[2], SARAH A. PENDERGRASS[1]

Biomedical & Translational Informatics, Geisinger Health System
100 N. Academy Ave. Danville, PA 17821, USA
Email: cbauer@geisinger.com

Department of Neurological Sciences, University of Vermont College of Medicine
149 Beaumont Ave. Burlington, VT 05405, USA

The past decade has seen exponential growth in the numbers of sequenced and genotyped individuals and a corresponding increase in our ability of collect and catalogue phenotypic data for use in the clinic. We now face the challenge of integrating these diverse data in new ways new that can provide useful diagnostics and precise medical interventions for individual patients. One of the first steps in this process is to accurately map the phenotypic consequences of the genetic variation in human populations. The most common approach for this is the genome wide association study (GWAS). While this technique is relatively simple to implement for a given phenotype, *the choice of how to define a phenotype is critical*. It is becoming increasingly common for each individual in a GWAS cohort to have a large profile of quantitative measures. The standard approach is to test for associations with one measure at a time; however, there are many justifiable ways to define a set of phenotypes, and the genetic associations that are revealed will vary based on these definitions. Some phenotypes may only show a significant genetic association signal when considered together, such as through principle components analysis (PCA). Combining correlated measures may increase the power to detect association by reducing the noise present in individual variables and reduce the multiple hypothesis testing burden. Here we show that PCA and k-means clustering are two complimentary methods for identifying novel genotype-phenotype relationships within a set of quantitative human traits derived from the Geisinger Health System electronic health record (EHR). Using a diverse set of approaches for defining phenotype may yield more insights into the genetic architecture of complex traits and the findings presented here highlight a clear need for further investigation into other methods for defining the most relevant phenotypes in a set of variables. As the data of EHR continue to grow, addressing these issues will become increasingly important in our efforts to use genomic data effectively in medicine.

1. Introduction

In the past decade, genome wide association studies (GWAS) have revealed more than ten thousand associations between genetic loci and traits [1]. As GWAS continue to grow in number, sample size, and range of phenotypes, they offer an opportunity to begin to untangle the complex network underlying phenotypic variation. One challenge in this pursuit stems from an asymmetry in the genotype-phenotype map. While the range of genetic variation in humans is fairly well characterized and a given genome can be sequenced to arbitrary depth, there is no obvious way to measure a physiologically complete phenome or even outline how to divide it into separate units [2]. Even subtle choices in how a phenotype is defined can affect which loci associate with it [3, 4]. There is a growing need to analyze these choices and their effects if we wish to build a genotype-phenotype map that captures the relationships most relevant to biology and the clinic.

The first human GWAS defined phenotypes based on clinical case control status [5, 6, 7]. Binary phenotypes such as these are a natural choice if our ultimate goal is to predict disease risk, but diseases are typically diagnosed based on a number of underlying quantitative variables and expert opinions. For example, dozens of loci have been implicated in the risk of multiple sclerosis [8]. However, this condition is heterogeneous in its presentation and is diagnosed based on an accumulation of symptoms, quantitative measures, and subjective categorization, only after ruling out other conditions [9]. There are also subtypes of multiple sclerosis as well as other distinct but related demyelinating syndromes [10, 11]. This complexity makes it exceedingly difficult to understand how each of the associated gene variants might be contributing to the disease.

Recently we have begun to see association studies conducted in cohorts that have been given batteries of quantitative assays [12, 13] and comprehensive electronic health record (EHR) data is being used to construct phenotypic profiles. The availability of these large sets of traits has lead to an approach known as the phenome wide association study (PheWAS) where each variant is tested for associations with a range of phenotypes [14, 15]. Recent applications of PheWAS have revealed many novel genotype-phenotype associations and the potential for a large degree of pleiotropy within disease related traits [14, 16, 17]. Variants that associate with multiple traits could be indicative of genetic modules that underlie multiple diseases but in some cases they may simply represent partially redundant measures that correspond to a single disease state.

Given a profile of quantitative traits, multivariate techniques such as principal component analysis allow us to combine related variables into a set of statistically independent measures. Combining different raw measurements into new metrics can identify new associations that may provide important insights into the biology of complex traits and may provide better predictors of disease risk [18, 19]. Consider for example, four GWAS for height, weight, and body mass index (BMI), and type II diabetes. Even though BMI is simply a function of height and weight, the results of these three associations tests will not identify exactly the same sets of loci. Likewise, many variants associate with both BMI and type II diabetes, but a large part of this overlap stems from BMI being a risk factor for type II diabetes [20]. Metabolomic studies have also demonstrated that some gene variants show much stronger relationships with the the ratios of metabolites than they do with the absolute abundances of either molecule [18].

While an EHR can contain thousands of types of data, such as clinical laboratory measures, similar variables may be collected or reported in different ways. Logical observation identifiers names and codes (LOINC) provides unique numerical identifiers to distinguish relevant differences between laboratory measures [21]. Most analyses that have been conducted to date have involved laborious data harmonization procedures to ensure that grouped lab results measure the same quantity in the the same way [22]. With the large numbers and types of measures in the EHR, it is often not feasible to carefully harmonize each and every phenotype. Thus, it is important to explore approaches that will allow for high throughput use of multiple phenotypes.

Here, we have mined the Geisinger Health System EHR for quantitative measures to produce a high dimensional phenotypic profile for a large population of genotyped patients in the MyCode® Community Health Initiative. Using these data, we outline and compare three general strategies for identifying loci that associate with one or more components of this phenomic profile: PheWAS, PCA, and cluster PCA. Our results show that each of these methods can detect associations that are missed by the others and that the significance of a given association can vary by many orders of magnitude based on how a phenotype is defined. These findings set the stage for further use of EHR data in gene associations studies and highlight important considerations as we attempt to improve the predictive power of medical genomics and clinical phenotyping.

2. Methods

2.1. *Genetic Data*

All of the data described in this paper come from a cohort of patients in the MyCode Community Health Initiative at the Geisinger Health System. Each patient was genotyped for 659,010 SNPs with minor allele frequency greater than 1% using Illumina OMNI Express Exome chips. We excluded any SNPs that had call rates < 99%, sample call rates < 99%, as well as 113 SNPs that show large differences between batches. We restricted our analysis to individuals with greater than 99% likelihood of European ancestry, as defined by quadratic discriminant analysis using the first four principal components of ancestry based on the 1000 genomes project.

2.2. *Phenotypic Data*

For 38,269 patients in the Geisinger Health System that met these criteria, we extracted age, sex, BMI and the median values for the following 29 outpatient laboratory measures as defined by LOINC codes (Table 1). Most of the lab measures showed large deviations from normality at the population level, so we first performed Box-Cox transformations on each variable. Each variable was also centered and scaled by subtracting the mean value and dividing by the standard deviation.

2.3. *Imputation*

Within the set of lab data that we analyzed, 7.1% of patient-lab pairs had no results available. Nearly a third of the missing data came from ~6000, mostly young, individuals that lacked lipid measurements (Figure S1). We used predictive mean matching to impute all missing values. Imputation was performed in R, using the MICE package. Due to multicollinearity, within a subset of the 29 variables, we excluded 11 pairs of variables with correlation coefficients greater than 0.5 as predictors of each other. Aside from this restriction, each variable was modeled as a linear function of all other variable, include age, sex, and BMI. We performed 5 separate imputations, selecting among the 5 closest cases, over 120 iterations. Nearly all chains exhibited convergence with 20 iterations. In the majority of cases, the distribution of imputed values was indistinguishable from the original distribution (Figure S2).

Table 1. Definitions of the LOINC codes extracted from electronic health records.

LOINC	Description
718-7	Hemoglobin [Mass/volume] in Blood
4544-3	Hematocrit [Volume Fraction] of Blood by Automated count
787-2	Erythrocyte mean corpuscular volume [Entitic volume] by Automated count
786-4	Erythrocyte mean corpuscular hemoglobin concentration [Mass/volume] by Automated count
785-6	Erythrocyte mean corpuscular hemoglobin [Entitic mass] by Automated count
6690-2	Leukocytes [#/volume] in Blood by Automated count
789-8	Erythrocytes [#/volume] in Blood by Automated count
788-0	Erythrocyte distribution width [Ratio] by Automated count
32623-1	Platelet mean volume [Entitic volume] in Blood by Automated count
777-3	Platelets [#/volume] in Blood by Automated count
2345-7	Glucose [Mass/volume] in Serum or Plasma
2160-0	Creatinine [Mass/volume] in Serum or Plasma
2823-3	Potassium [Moles/volume] in Serum or Plasma
3094-0	Urea nitrogen [Mass/volume] in Serum or Plasma
2951-2	Sodium [Moles/volume] in Serum or Plasma
2075-0	Chloride [Moles/volume] in Serum or Plasma
2028-9	Carbon dioxide, total [Moles/volume] in Serum or Plasma
17861-6	Calcium [Mass/volume] in Serum or Plasma
1743-4	Alanine aminotransferase [Enzymatic activity/volume] in Serum or Plasma by With P-5'-P
30239-8	Aspartate aminotransferase [Enzymatic activity/volume] in Serum or Plasma by With P-5'-P
1975-2	Bilirubin.total [Mass/volume] in Serum or Plasma
2885-2	Protein [Mass/volume] in Serum or Plasma
10466-1	Anion gap 3 in Serum or Plasma
751-8	Neutrophils [#/volume] in Blood by Automated count
2093-3	Cholesterol [Mass/volume] in Serum or Plasma
2571-8	Triglyceride [Mass/volume] in Serum or Plasma
2085-9	Cholesterol in HDL [Mass/volume] in Serum or Plasma
13457-7	Cholesterol in LDL [Mass/volume] in Serum or Plasma by calculation
2965-2	Specific gravity of Urine

2.4. *Principal Component Analysis*

For each imputed dataset, we performed principal component analysis (PCA) in R, using the *prcomp* function. The PCA results were nearly identical within each imputed dataset. The average angle between all ordered pairs of Eigenvectors for the first 19 components was 4.9° and the only angles greater than 20° were caused by an alternation in the order of components 20 and 21 in some of the analyses (Figure S3). Given the minimal differences between the imputed data sets, we chose the first imputed data set to use in all downstream analyses.

2.5. *K-means clustering*

Using K-means clustering, we divided our 29 variables into 7 clusters based on their pairwise absolute correlations (Figure 1). The distance between two LOINC codes was defined as $1\text{-}R^2$. Clustering was performed in R using the *kmeans* function with 200 random starting clusters. Since sum of squares measures did not indicate an optimal number of clusters, we choose the maximum number of clusters where all clusters contained at least 3 phenotypes.

2.6. *GWAS*

We first performed associations between all 29 phenotypes individually (Figure S4). We also performed associations with 29 principle component scores (Figure S5). Finally, we performed associations with scores of the principal components within each cluster (Figure S6). All association tests were performed using PLATO 2.0 (https://ritchielab.psu.edu/plato). In each case, we modelled the principal component score as an additive function of allele count with age, bmi, sex, and the first four principal components of ancestry included in the model as covariates.

3. Results

Our phenotypic dataset comprised 29 outpatient clinical lab measures extracted from Geisinger Health System EHR. In order to ensure compatibility with other datasets, we choose to include only measures that complied with the LOINC standard of medical laboratory observations [23]. For each of the 29 clinical lab measures, we performed a separate GWAS in PLATO. Using these measures, we identified 6361 statistically significant associations (FDR < 0.01). Every lab measure had multiple SNPs associated with it, ranging from 12 SNPs for chloride concentration in blood to 783 SNPs for the number of leukocytes per unit of blood (Figure 1). Of these associations, 31% involved a SNP that was linked to more than one lab measure.

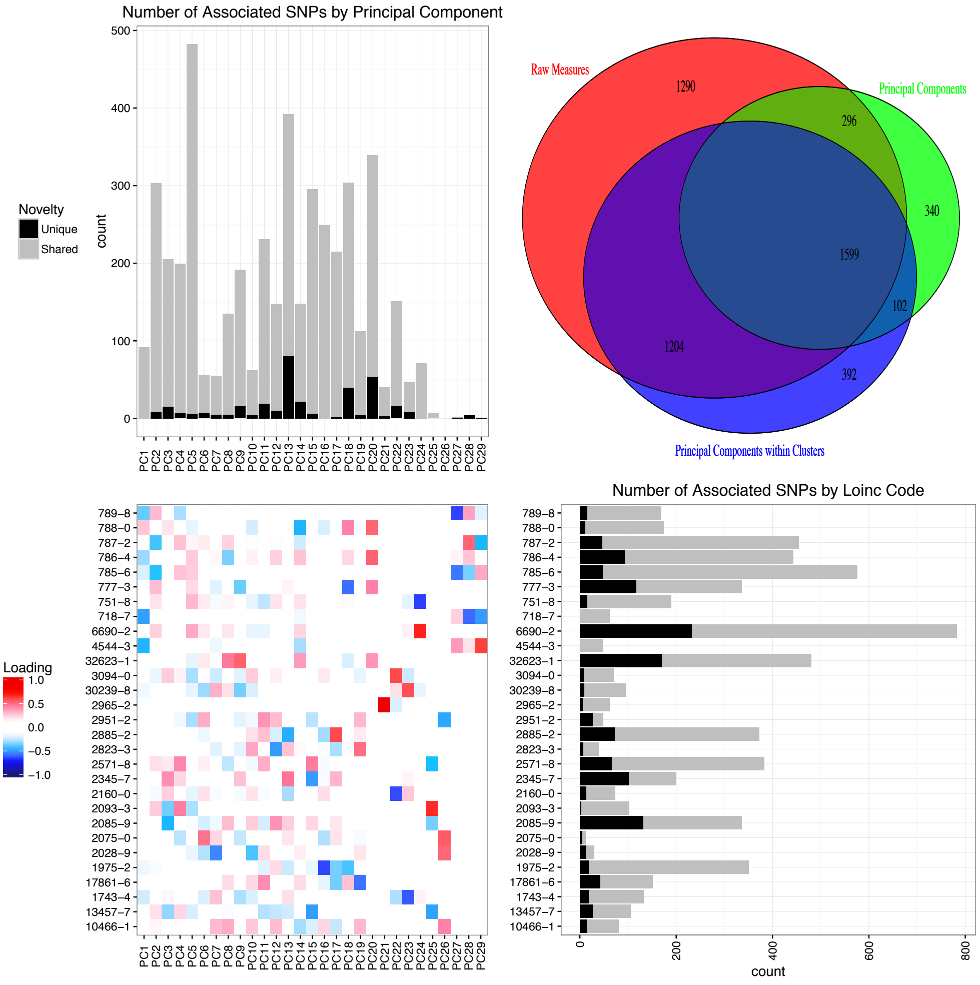

Figure 1. Associations detected with LOINC measures, PCA, and cluster PCA. The Venn diagram in the upper right panel shows the number of unique and shared SNPs that were associated with a phenotypic measure as defined by each of the three methods. The upper left panel shows the number of SNPs that associated with each principal component. The lower right panel shows how the associations were distributed across the LOINC measures. Gray bars represent the total number of SNPs while black shows the number that are unique to that measure. The bottom left panel shows how each of the phenotypes defined by LOINC codes loads onto each of the principal components.

Given that several groups of the lab results had very strong correlations and nearly all showed at least modest correlations with a few other variables (Figure 2) we hypothesized that statistical power might be improved by combining highly correlated measures. To test this, we performed principal component analysis on the combined set of all 29 lab measures. A plot of the cumulative variance explained by each additional component was smooth and increased gradually indicating that even the highest components might be measuring physiologically meaningful traits (Figure S7).

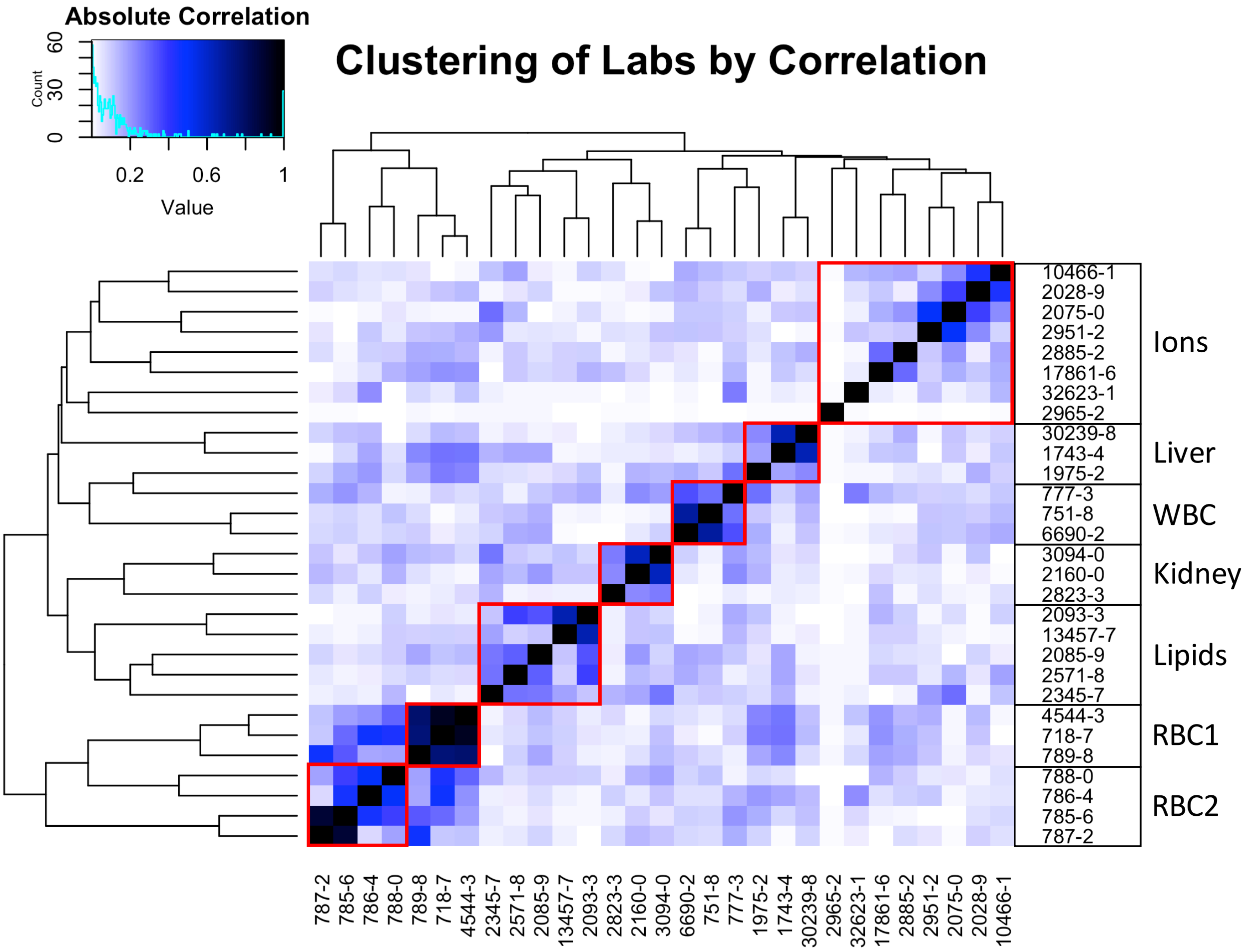

Figure 2.
Clustering of LOINC measures into related groups. The heat map indicates the absolute value of the correlation coefficient between all pairs of LOINC codes. Each cluster, as defined by k-means, is indicated by a red bounding box. The names on right column indicated the functional categories that describe each cluster.

We next performed GWAS for all 29 principal components, just as we did for the original measures (Figure S5). This analysis resulted in 4536 significant associations (FDR = 0.01). We expected to see a reduction in the total number of significant associations as one principal component could capture variation from multiple raw measures. Surprisingly, 48% of these associations involved a SNP that was linked to multiple components. Figure 1 shows a Venn diagram comparing the number of unique and overlapping associations across the various approaches for phenotypes used in this paper. Although 2494 of the SNPs that associated with one or more of the LOINC measures did not show a significant association with any of the principal components we did discover 442 new associations using these scores. 1895 SNPs associated with both a raw measure as well as a principle component (PC).

PC5 had the largest number of significant associations, 482, followed by PC13 with 392 and PC20 with 339. There was no clear pattern in how the significant associations were distributed among the first 24 components, although there were practically no associations with PC25-29 (Figure 1). In PCA, the first few components often capture a large percentage of the variation so it was interesting to see so many SNPs associating with higher components while PC1 only had 92 associated SNPs. Further analyses provide some insights. First, if we include age and BMI in the set of variables prior to PCA, we find that these variables load most strongly onto components 1. This makes sense given that age and BMI contribute to many physiological measures, especially among disease relevant traits. However, since these are both covariates in the GWAS regression model, it would be troubling to see many SNPs associated with PC1.

In PCA, the loadings indicate the magnitudes and directions that the original measures contribute to each component. Analyzing the loadings of some noteworthy components provided some additional clues to the causes of this behavior. PC2 was dominated by a few measures of blood cells: namely, the volume of erythrocytes moving in the opposite direction of cholesterol and the numbers of erythrocytes, white blood cells (WBCs), and platelets (Figure 1). PC5 was similar with WBC counts moving in the opposite directions of platelet volume and cholesterol. These associations may reveal overlap in genetic networks that regulate lipids and and the immune system. A number of studies have previously identified relationships between WBC counts and carotid plaque thickness, body fat percentage, and lipid profiles [24, 25, 26].

While PC2 and PC5 were linked to many loci, these were predominantly the same loci that were linked to one or more of the original measures. More relevant than the total number of associations detected is the number of associations that were unique to a principal component and not detectable using any of the original measures. PC13, PC18, and PC20 were responsible for the majority of these novel associations. PC13 measures a complex relationship among our measures in which serum levels of potassium and glucose vary inversely with total protein and creatinine. This is interesting because potassium and creatinine are highly correlated at the population level and both are diagnostic of kidney function. 89% percent of the associations with PC13 also map to the HLA locus suggesting a relationship between the adaptive immune system and these blood measures. PC18 and PC20 both measure relationships among erythrocyte distribution width, hemoglobin, and platelet measures (Figure 1).

Overall, the principal component approach detected fewer total significant associations than the LOINC measures, but a few components did allow us to identify novel associations. The principal components that proved most useful in this regard seemed to load primarily off of 2-6 measures (Figure 1) and those measures tended to be closely related. Components that were dominated by a single measure or had large number of weak loadings did not yield many novel results. These observations suggested a third approach. If we first divided the original 29 measures into small groups of related traits before performing PCA, we might restrict our range of phenotypes to space that corresponds better to the ways that gene variants actually impact phenotype.

Using K-means clustering, we divided our 29 variables in 7 clusters based on their pairwise absolute correlations. The choice of the number of clusters was somewhat arbitrary as the sums of square both within and between clusters never reached obvious plateaus. The choice of 7 clusters resulted in each group containing 3-8 measures, which corresponded well to our desired range, and it also broke them into groups that made intuitive sense (Figure 2). For example, all of the white blood cell counts formed a single cluster, and all of the lipid measures clustered together with serum glucose. We then performed PCA within each of these clusters and used these principal component scores to run a third GWAS with the same parameters as the previous two (Figure S6).

The genetic variants that associated with the scores of the cluster principal components had much larger overlap with original measures, sharing 2803 SNPs, but it also revealed 392 new SNPs that did not associate with either the original measures or the principal components of the entire data set (Figure 1). The distribution of these new associations varied greatly among each cluster (Figure S8). Within the ions cluster, the majority of the SNPs showed stronger associations with one of original measures than they did with any principal component (Figure S14). Within the three phenotypes that compose the liver cluster (1743-4: alanine aminotransferase, 30239-8: aspartate aminotransferase, and 1975-2: bilirubin), the associations detected for all three principal components correlated almost perfectly with those of one of the original measures (Figure S15). However, within the red blood cell cluster 1 (718-7: hemoglobin, 4544-3: hematocrit, and 789-8: erythrocytes), nearly all of the alleles tested showed their strongest association with one of the principal components (Figure 3).

Within each cluster, the middle components were the most likely to have novel associations. In general, PC1 had associations that were very similar in their significance levels to those found with the original measures. With each successive PC, the p-values would usually become more significant with respect to the LOINC measures, but less significant in absolute terms due to the reduction in total variance with each PC. In the red blood cell 1 (RBC1) cluster, nearly all of the novel significant associations occur with PC2 (Figure 3). A high score in this component corresponds to a low count of erythrocytes per unit volume of blood, but a high hematocrit score, and hemoglobin concentration. Since none of these associations were not found using erythrocyte mean corpuscular volume (787-2) as the phenotype of interest, it seems that there are a large number of gene variants linked to the concentration of hemoglobin within erythrocytes. A Manhattan plot shows that these new associations come from many distinct loci (Figure 4).

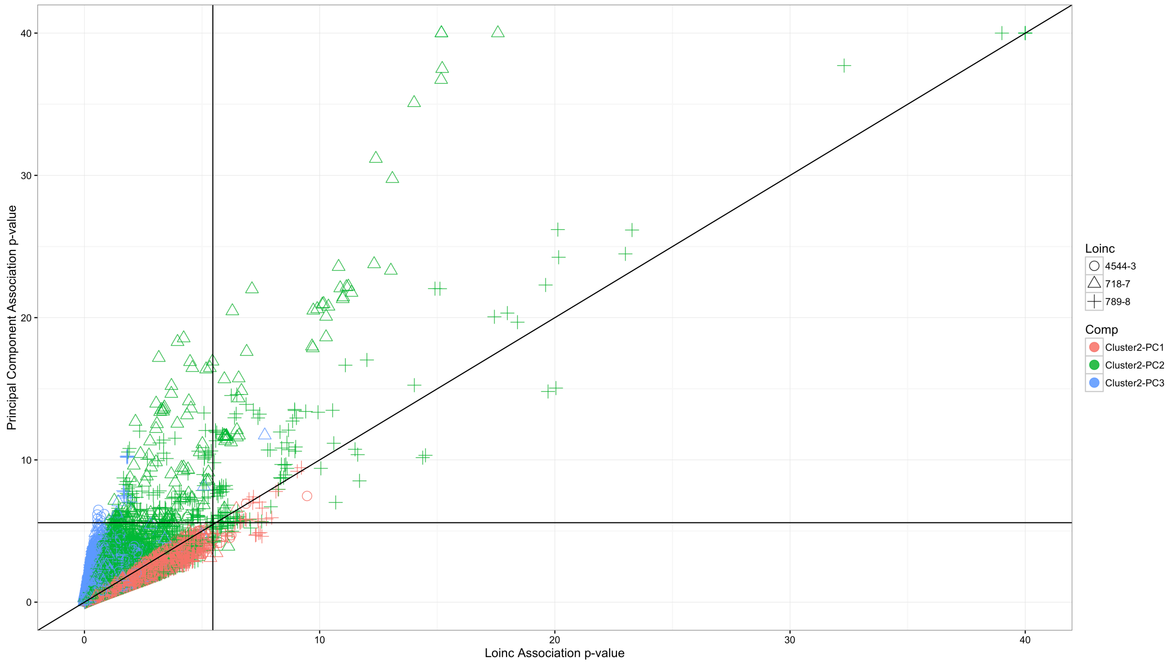

Figure 3.
Comparison of p-values for associations with the principal components and LOINC measures that compose the red blood cell cluster 1. Each point in the scatter plot represent one SNP. Both axes are scaled to the negative log base ten of the p-values. The y-axis indicates the lowest p-value that a given SNP had with any of the principal components. The components are coded by the color or the point. The x-axis indicates the lowest p-value that a given SNP had with any of the LOINC measures. The LOINC measures are coded by the shape of the point.

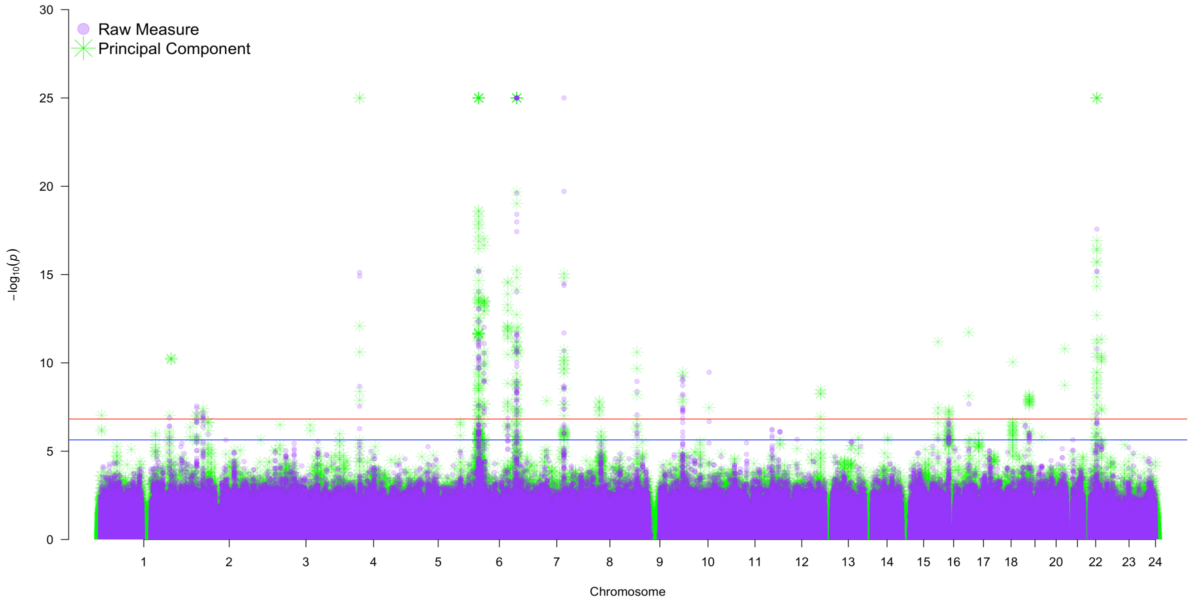

Figure 4.
Manhattan plot of the associations detected for the RBC1 cluster. The x-axis indicates the chromosomal coordinate and the y-axis shows the negative log base ten of the association p-value. Associations with any of principal components and LOINC measures are displayed in green and purple, respectively. The red line indicates a false discovery rate of 0.001 and the blue line indicates a false discovery rate of 0.01.

4. Discussion

Our results demonstrate that the choice of how to define a phenotype can have a large impact on our ability to detect relationships with genetic loci. Given a set of quantitative trait measures, we have outlined three different strategies for defining phenotypes prior to association testing. The standard method is to simply test against whatever phenotypic measures are in hand, without any additional considerations. While some measures of phenotype may be arbitrary or based purely on convenience, this may still be the most reasonable choice in many situations. In this particular case, the original phenotypes come from clinical lab tests that are prescribed because they have proven to be useful diagnostics and we find the greatest number of significant associations using these measures alone.

In spite of this generality, our results also indicate that many genotype-phenotype connections are not apparent when phenotypes are considered individually. Using two different methods based on principal component analysis, we have increased the number of significant associations that we could detect by 19%. Given the extremely large number of hypotheses that are tested in a single GWAS experiment, the p-value threshold for significance must be correspondingly low. Most segregating alleles have relatively small impacts on any given phenotype and we are unlikely to detect a significant association unless the phenotype of interest aligns very well with the effect of the variant. The majority of true positive results will inevitably fail to reach the significance threshold.

Principal component analysis provides one strategy for overcoming some of these obstacles. When performed on the entire dataset, it has the ability to capture relationships between diverse phenotypic measures. In this case, the components with the largest variance did not provide much new information. This is likely because these components capture the covariance of large numbers of measures that relate to the biggest sources of phenotypic variation in populations, such as age. There should not be genetic determinants of age, except in extreme cases, and even if there were, it is common practice to control for the effects of age in regressions.

The greatest utility of this method comes from the middle order components that capture more complex relationships. In our analyses, PC13 was related to a complex interaction between serum concentrations of potassium, creatinine, glucose, and total protein. It is not yet clear how this relates to human physiology, but the fact that 79 SNPs, distributed widely across the HLA locus, associate more strongly with this principal component than they do with any of the measures that contribute to it suggests some underlying mechanistic connection between this combination of variables and the function of the immune system. Perhaps phenotypic profiles such as this will also prove to be useful indicators of disease risk or progression.

This approach also allows us to observe effects that are orthogonal to the primary axis of variation. For example, creatinine and urea levels are both indicative of kidney function and they are very highly correlated at the population level. However, urea is a byproduct of all protein metabolism while creatinine is produced only by muscles so it is reasonable to assume that various genes could influence these traits independently. Indeed, principal component 22 corresponds to an inverse relationship between these two variables and several variants associate with the ratio of creatinine to urea in the blood without a detectable relationship to either variable in isolation.

One of the weaknesses of using principal component scores from a large dataset is that the eigenvectors correspond to the maximal variance within a set of measures which may not have any relationship to how traits are influenced at the gene regulatory level. A SNP that might correspond to elevated total cholesterol is unlikely to affect every other trait that correlates with cholesterol in a population. It can also become difficult to extract meaning from a principal component that is influenced by many, potentially disparate measures. If we hope to translate research findings into clinically relevant information, it can be useful to limit our search space to a number of dimensions that a human can understand. In order to strike balance between exploring the full range of complex interactions in biology and maintaining the ability to interpret our results, we also investigated a third approach that involved clustering our data based on the correlation structure of the variables prior to performing PCA.

While this did not improve our power in all cases, several groups of related measures yielded many more genetic associations, and at least a few new associations were discovered within each cluster. In particular, assays of blood cells and kidney function seem to benefit the most from this technique. The first 3 components of the RBC2 cluster collectively associate with 134 SNPs that do not show significant associations with any other measure that we tested. These components each measure different ways that the variance in erythrocyte size relates to hemoglobin concentration and mean erythrocyte volume. It is interesting to note that PC3 from this cluster had the most unique associations and is related to PC20 from the global PCA, which also identified new SNPs. Within the kidney cluster, PC3 measures the difference between urea and creatinine levels and associates with 41 unique variants. Again, this is related to PC22 from the global analysis. The fact that both clustered and global PCA identify associations with complex interactions between multiple blood cell and kidney function measurements indicates that the genetic regulation of these traits is not captured well by any single measure. It will be interesting to test if these same interactions are linked to the prevalence or prognosis for any disease states.

It is likely that numerous other combinations of the underlying measures would yield even more connections between gene variants and phenotypes but there is no way to exhaustively explore them. As the number of phenotypic measures that we can collect for a GWAS cohort continues to grow, it will be increasingly important to develop better strategies for specifying exactly which measures to choose test for associations. Further investigation into this topic will be critical to gaining insight into gene function and has deep implications for how we think about concepts such as pleiotropy.

Acknowledgements

We would like to acknowledge Marylyn Ritchie, Marta Byrska-Bishop, Anna Basile, Anurag Verma, and H. Lester Kirchner for helpful discussions.

References

1. D. Welter et al., *Nucleic Acids Res.* **1;42**, 1001 (2014).
2. M. Samuels, *Curr Genomics*. **11(7)**, 482 (2010).
3. W. Bush and J. Moore, *PLoS Comput Biol*. **8(12)** (2012).
4. E. Stergiakouli et al., *Obesity.* **10**, 2252 (2014).
5. R. Klein et al., *Science*. **308**, 385 (2005).
6. A. Dewan et al., *Science*. **314**, 989 (2006).
7. The Wellcome Trust Case Control Consortium, *Nature*. **447**, 661 (2007).
8. International Multiple Sclerosis Genetics Consortium et al., *Nat Genet*. **45(11)**, 1353 (2013).
9. S. Katz, *Curr Opin Neurol*. **28(3)**, 193 (2015).
10. T. Avsar et al., *PLoS One*. **5;10(5)**, e0122045 (2015).
11. D. Karussis, *J Autoimmun*. **48-49**, 134 (2014).
12. Y. Kamatani et al., *Nat Genet.* **42(3)**, 210 (2010).
13. K. Suhre et al., *Nat Genet.* **43(6)**, 565 (2011).
14. J. Denny et al., *Nat Biotechnol*. **31**, 1102 (2013).
15. S. Pendergrass et al., *Hum Hered.* **79(3-4)**, 111 (2015).
16. S. Pendergrass et al., *PLoS Genet*. **9(1)**, e1003087 (2013).
17. M. Hall et al., *PLoS Genet*. **4;10(12)**, e1004678 (2014).
18. C. Geiger et al., *PLoS Genet*. **4(11)**, e1000282 (2008).
19. E. Stergiakouli et al., *Obesity*. **22(10)**, 2252 (2014).
20. P. Visscher et al., *Am J Hum Genet*. **13;90(1)**, 7 (2012).
21. A. Forrey et al., *Clin Chem*. **42(1)**, 81 (1996).
22. S Bennett et al., *Genet Epidemiol*. **35(3)**, 159 (2011).
23. J. Deckard et al., *J Am Med Inform Assoc.* **22(3)**, 621 (2015).
24. S. Mitchell et al., *Stroke.* **32**, 842 (2001).
25. M. Farhangi et al., *J Health Popul Nutr.* **31(1)**, 58 (2013).
26. L. Ferreira et al., *Rev. Bras. Hematol. Hemoter*. **35**, 3 (2013).