Characterizing genomic and functional variation among populations and tissues

Margaret Antonio and Nicole Ferraro

GENE 245 Final Report

Abstract

Considerable effort has been directed at elucidating functional variation in the human genome to understand the genetics underlying changes in cellular activities. However, variants can differ greatly in frequency amongst populations, and also have differential effects in tissues. Here, we aim to understand how we can leverage genomic variation among populations to infer sample characteristics, and how genomic variation intersects with functional variation. For pipeline development, we look at variation in a gene that varies between populations, ABCA12, and two different tissues, skin and brain, for functional annotations. We find that a single variant in ABCA12 with high change in derived allele frequency for African and European populations or a subset of over 100 randomly selected variants is sufficient to classify an individual's ancestry with over 90% accuracy. Furthermore, we find that genomic variation is depleted in functional regulatory regions as compared to the background genome. We anticipate that the methods described here can be extended to additional areas of variation in the genome, and used to connect that variation to other phenotypes, such as disease.

Introduction

Around 10 million single nucleotide polymorphisms (SNPs) can be found in the human genome, with any individual carrying about 3 million SNPs. SNPs are positions in the genome that vary between individuals, and to which much of the phenotypic difference observed among humans is attributed. Furthermore, certain SNPs or sets of SNPs are unique to individuals of a given population, described by ancestry or phenotype (e.g. a genetic disease). The 1000 Genomes Project, Human Reference Consortium, and others have made significant progress in assembling and annotating a globally representative collection of SNPs and their frequencies in namely ancestral populations [4,8].

Given this vast amount of data (around 2504 individuals, 26 populations, and over 88 million variants in The 1000 Genomes Project alone), one could probe the predictive power of variants in classifying populations. Here, we explore this possibility in ancestral populations, however, one could apply this to any type of population (e.g. in disease). Single variants have been shown to have large differences in derived allele frequency (DAF) between populations. For example, the SNP rs10180970 in ABCA12, a gene involved in skin pigmentation and permeability, has a DAF of 0.96 in the Asian, 0.91 in European, and 0.13 in African populations [2]. Similarly, an indel in ZNF804A has a DAF of 0.83 in the Asian population versus 0.58 and 0.07 in European and African populations, respectively [2]. These population specific variants are suspected to have faced regional specific selection, such as pathogen resistance in China in the case of ZNF804A. Here, we explore the predictive power of these highly δDAF single SNPs compared to random sets of SNPs in a given gene and across a whole chromosome for population classification as a first step to studying variation in differentiating populations.

As a second step in exploring variation, we look at functional variation in tissues as cell populations. Functional variation is essential to the ability of cells to perform different roles, despite identical genomic information. Previously, it has been shown that there is a higher level of conservation for genetic sequences associated with epigenetic modifications, specifically surrounding nucleosomes [3]. Since genes will have different contextual roles, quantifying the variants that intersect functional regions can provide insight as to the impact of that functional region on a gene's performance, and potential population differences.

Methods and Results

Population variation

To investigate the role of SNPs in distinguishing populations, we first performed K-Nearest-Neighbors (KNN) for classification based on subsets of variants and then principle component analysis (PCA) for visualization of all variants for 1164 individuals (503 European and 661 African) from the 1000 Genomes Project, Phase 3 [4]. For this analysis, we focused on variants in chromosome 2 and then on variants in ABCA12 and ZNF804A. Variant sites were filtered to contain only bi-allelic sites and sites without missing information for any individual. The final set included 5831 bi-allelic variants. The phased genotypes for individuals were recoded to numeric genotypes where 0 if homozygous for the reference alleles, 1 if heterozygous for the alternate allele, and 2 if homozygous for the alternate allele.

We hypothesized that given previous findings of high differentiation for populations in ABCA12, variants in this gene would be the best predictors in population classification. We performed KNN with 10-fold cross validation, repeated three times, with a 10% hold-out on the genotypes for all individuals using the caret R package [9]. We used area under the ROC curve (AUC) to determine an optimal number of neighbors (k), 43. Using all variant sites in ABCA12 as features for classification produced an AUC of 0.98. To test if this high prediction accuracy was a function of the variants in ABCA12 or the number of variants, we repeated this procedure with sets of 5 to 6000 random variants from chromosome 2 and from ABCA12. Results are shown in Fig. 1 for chromosome 2 (panel A) and ABCA12 (panel B) with the ROC for the variant, rs10180970, shown in a separate study to have a high δDAF between the two populations [2]. The AUCs for the random chromosome-wide variant subsets match closely to those of the subsets in ABCA12. Using rs10180970 as a single feature shows better performance than random chromosome-wide and ABCA12 variant sets of less than 75, but worse than sets of over 100 variants.

We used PCA to qualitatively assess separation between the populations. As shown, 6000 randomly sampled variants in chromosome 2 enables clear separation of the two populations (Fig.1C), while separation is less clear using all 5831 variants in ABCA12 (Fig. 1D). These results are consistent with the high prediction accuracy using large samples sizes of random variants in chromosome 2 and ABCA12 (Fig. 1 AB). This analysis was performed for ZNF804A gene-wide variants and two high δDAF variant sites in ZNF804A for African and East Asian populations. Results were very similar with comparably high prediction

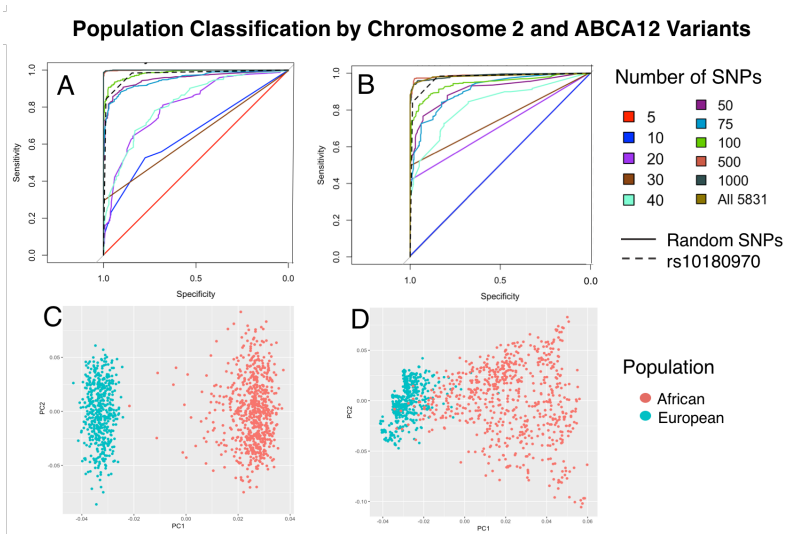

Figure 1: KNN with 10-fold cross validation on 1164 individuals with 10% hold-out. ROC curves of KNN classification for random SNPs in chromosome 2 (A) and in ABCA12 (B) compared to single variant rs10180970. Separation of populations by PCA for 6000 random variants from chromosome 2 (C) and 5831 variants in ABCA12 (D).

accuracies (∼98%) using just the high δDAF variants as features and using random sets of over 100 variant loci.

Functional annotations

After exploring ABCA12 variants' implications in differentiating populations, we sought to understand the functional annotations possibly impacting the genes role. We looked at DNase-sequencing data, from the ENCODE consortium, and hypothesized that ABCA12, with its differential role in human populations, would have an over-representation of regulatory regions. Previous work demonstrated that a motif in the upstream region of ABCA12 is essential for promoter function [5], and this was a motivating factor in examining ABCA12 functional annotations. Because ABCA12 is involved in skin pigmentation, we collected skin tissue samples, 16 total. We used post-processed data, so peak calling and filtering had already been completed. We subset to chromosome 2 for all analyses. To assess if DNase peaks were enriched in the region of ABCA12, we created a background distribution of the number of peaks in identically sized regions elsewhere in chromosome 2. One thousand random regions were selected, and peaks were counted within those regions. A negative binomial was fit to this distribution, with the fitdistrplus R package [6].

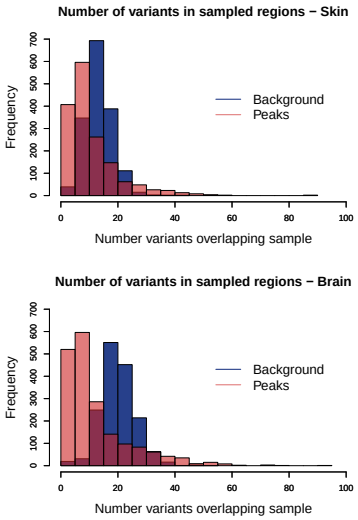

Figure 2: **Variant-peak overlap in brain and skin.**

We calculated the probability that the number of counts seen in the region of interest was greater than expected based on the model. None of the samples showed a significant enrichment of peaks in ABCA12, defined by a probability greater than 0.95. They ranged from 0.445 to 0.741 for 16 samples. After assessing the functional enrichment surrounding

ABCA12 in skin samples, we examined a different tissue, with a similar sample size, brain, and repeated the analysis for 19 samples. No samples showed significant peak enrichment, though the probabilities were lower than in skin, and had a wider range, from 0.265 to 0.864.

Previous work showed that regulatory regions are depleted of indel mutations, but SNPs were enriched, particularly around bulk nucleosomes [3]. We intersected variants from 1000 Genomes with DNase-seq peaks and compared to a background distribution. We randomly sampled 100 peak regions from each sample and collapsed across, calculated the number of variants, and then sampled the number of variants in an equal number of random regions. We determined if the mean of these two distributions significantly differed via a t-test with unequal variances. The two distributions are seen in Fig. 2. The number of variants in peak regions was significantly less than in the background, with $p<2.2e-16$ in both tissues, demonstrating the depletion of variants in peak regions for this dataset. We then looked at the number of variants intersected with regulatory regions in ABCA12, but no variant-peak overlaps occurred in these regions. The analysis was repeated in ZNF804A, and the trends were repeated, though two samples had significant peak enrichment in skin.

Conclusions and Future Work

Population classification by variants showed that single high δDAF SNPs have high predictive power for two population classification, but are matched in prediction ability by randomly chosen sets of >100 SNPs. Although we expected low predictive power for random sets of SNPs, this result is consistent with a study which found that (1) the number of loci was the most important variable in population classification and (2) that 100 loci were sufficient to classify populations with near 100% accuracy [6]. This is also consistent with the clear population separation by PCA for 6000 random variant loci. However, the lack of as clear separation of populations for around the same number of variants in ABCA12 warrants further study. Overall, the results point to the clear presence of global genomic variation between populations that can be captured by a set of >100 loci. This is an important consideration in cross-population GWAS, where it is important to include a large panel to capture population differences.

We extended our analysis of genomic variation to understand variant intersection with functional annotations. No samples showed regulatory region enrichment of in ABCA12, and only two in ZNF804A, so no clear relationship emerged between function and number of regulatory regions. We then looked at whether the number of variants present in regulatory regions was enriched or depleted compared to the background, and found a significant depletion in variation in regulatory regions, consistent with previous work [3]. This was not consistent across all types of variation in prior work though, and so indicates an area of further investigation. Our analysis was restricted to chromosome 2, so in the future we want to explore the overlap of genomic and functional variation across the genome. Overall, this analysis explores the relation of genomic variation to phenotype, and the intersection between genomic and functional annotations, and presents a potential pipeline that could extend to additional genomic regions and populations.

References

1. Dilthey, A., et al (2015). Improved genome inference in the MHC using a population reference graph. Nat. gen, 47(6), 682-688.
2. Colonna, V., et al (2014). Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome biol, 15(6), R88.
3. Tolstorukov, M. Y., et al. (2011). Impact of chromatin structure on sequence variability in the human genome. Nat. struct. mol. Biol., 18(4), 510-515.
4. Consortium, T. 1000 G. P. A global reference for human genetic variation. Nature 526, 6874 (2015).
5. Shimizu, Y., Ogawa, Y., Sugiura, K., Takeda, J. I., Sakai-Sawada, K., Yanagi, T., ... & Akiyama, M. (2014). A palindromic motif in the 2084 to 2078 upstream region is essential for ABCA12 promoter function in cultured human keratinocytes. Scientific reports, 4.
6. Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. Journal of Statistical Software, 64(4), 1-34.
7. Witherspoon, D. J. et al. Genetic Similarities Within and Between Human Populations. Genetics 176.1 (2007): 351359. PMC. Web. 5 June 2017.
8. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48, 12791283 (2016).
9. Kuhn, Max. "Building Predictive Models in R Using the caret Package." Journal of Statistical Software [Online], 28.5 (2008): 1 - 26. Web. 6 Jun. 2017