

Setser, Casandra H., Ancestry Informative Markers Tailored to Hispanic Populations. Doctor of Philosophy (Biomedical Sciences), January 2020, 117 pp., 16 tables, 10 figures, 136 references.

### **Abstract**

Hispanic populations are highly heterogeneous despite being grouped together as a conglomerate population; this makes an accurate panel of ancestry informative markers (AIMs) especially important for human identification.

In Chapter 2, the Genomic Origins and Admixture in Latinos (GOAL) dataset containing 494,886 SNPs was used for SNP ascertainment. Utilizing a country attributable variant of Wright's  $F_{ST}$ , 234 SNPs were selected for biogeographic ancestry (BGA) determination by tailoring each SNP to genetic differentiation of specific populations. Accuracy of BGA prediction was tested using multinomial logistic regression (MLR) and as few as 55 SNPs were robust to 90% for all populations studied. The panel of 234 SNPs was compressed by 65.8% to 80 SNPs by decreasing the influence of Honduras and the Dominican Republic SNPs with high country attributable mean  $F_{ST}$  values in favor of additional SNPs for Colombia, Cuba, and Puerto Rico; this balanced small panel size with classification accuracy.

In Chapter 3, the Setser80 Hispanic AIMs panel was tested against the panels of 128 SNPs developed by the Seldin group and 55 SNPs developed by the Kidd group using STRUCTURE, PCA, a naïve Bayesian classifier and MLR. In STRUCTURE, the Setser80 was able to distinguish Honduras, the Dominican Republic, and Colombia at  $K=4$ , where the Seldin and Kidd panels were optimized at  $K=3$  and distinguished only Honduras and the Dominican Republic; similar results were obtained by PCA. The GOAL dataset was combined with the Admixed American super-population from the 1000 Genomes Project to test the panel on an expanded dataset of seven populations. Overall, the Setser80 had superior results to the Seldin and Kidd panels with 91.5% accuracy by naïve Bayesian classifier and 93.2% by MLR. As an indication of its portability, the Setser80 had accuracies of >98% for Peru and >80% for Mexicans living in Los Angeles, which were not involved in SNP ascertainment. Given its accuracy and lack of overlap, the Setser80 may supplement existing panels for more granular Hispanic BGA determination.

In Chapter 4, the application of allele frequencies to forensic genetics, genealogy, and clinical genetics are discussed as well as future directions and ethical considerations.

ANCESTRY INFORMATIVE MARKERS

TAILORED TO HISPANIC

POPULATIONS

Casandra Hernandez Setser, B.S., M.S.F.S.

APPROVED:

---

Deanna Cross, Ph.D, Major Professor

---

John Planz, Ph.D., Committee Member

---

Robert Barber, Ph.D., Committee Member

---

Nicole Phillips, Ph.D., Committee Member

---

Raghu Krishnamoorthy, Ph.D., University Member

---

Bruce Budowle, Ph.D., Interim Department Chair

---

J. Michael Mathis, Ph.D., Ed.D., Dean  
Graduate School of Biomedical Sciences

**ANCESTRY INFORMATIVE MARKERS  
TAILORED TO HISPANIC  
POPULATIONS**

DISSERTATION

Presented to the Graduate Council of the  
Graduate School of Biomedical Sciences

University of North Texas

Health Science Center at Fort Worth

In Partial Fulfillment of the Requirements

For the Degree of  
DOCTOR OF PHILOSOPHY

By

Cassandra Hernandez Setser, M.S.F.S.  
Fort Worth, Texas  
January 23, 2020

## ACKNOWLEDGEMENTS

Family: Mom & Dad, Greg, Liliana;  
Grandmothers Olga Manzano, Gloria Manzano & Delia Hernandez

Friends: Ricardo Belmares, Jessica Juarez, Michael Nolan, Parul Chaudary, Elizabeth Nelson,  
Elizabeth Kerl, Gita Pathak, Frank Wendt, Laura Gaydosh-Combs

Professors: Arthur Eisenberg, Ranajit Chakraborty, Deanna Cross, John Planz, Robert Barber,  
Nicole Phillips, Raghu Krishnamoorthy, Rhonda Roby

## TABLE OF CONTENTS

|                                                                        |     |
|------------------------------------------------------------------------|-----|
| ACKNOWLEDGEMENTS .....                                                 | ii  |
| TABLE OF CONTENTS .....                                                | iii |
| LIST OF TABLES .....                                                   | vii |
| LIST OF FIGURES .....                                                  | ix  |
| <br>                                                                   |     |
| CHAPTER 1: Introduction and Background .....                           | 1   |
| 1.1. Migration and Admixture .....                                     | 3   |
| 1.1.2. From East Asians to Indigenous Peoples .....                    | 3   |
| 1.1.3. Hispanic Admixture Complexity .....                             | 4   |
| 1.2. Project Design & Chapter Summary .....                            | 7   |
| 1.3. Instruments of Ancestry Determination .....                       | 9   |
| 1.3.1. Ancestry Informative Markers .....                              | 9   |
| 1.3.2. Biostatistics & Bioinformatics .....                            | 11  |
| 1.3.2.1. Linkage Disequilibrium (LD) .....                             | 11  |
| 1.3.2.2. $F_{ST}$ for Population Differentiation .....                 | 12  |
| 1.3.2.3. Clustering Algorithms .....                                   | 14  |
| 1.3.2.4. Modeling & BGA Classification via Snipper 2.5 App Suite ..... | 14  |
| 1.4. Conclusions .....                                                 | 15  |
| 1.5. References .....                                                  | 16  |

|                                                                                                                                      |    |
|--------------------------------------------------------------------------------------------------------------------------------------|----|
| CHAPTER 2: SNP Ascertainment for Heterogeneous Hispanic Populations using the Genomic Origins and Admixture in Latinos Dataset ..... | 21 |
| 2.1. Introduction .....                                                                                                              | 22 |
| 2.2. Materials and Methods .....                                                                                                     | 24 |
| 2.2.1. The Genomic Origins and Admixture in Latinos (GOAL) Study .....                                                               | 24 |
| 2.2.2. SNP Ascertainment .....                                                                                                       | 24 |
| 2.2.3. Overview of $F_{ST}$ .....                                                                                                    | 25 |
| 2.2.4. AIMs Panel Creation .....                                                                                                     | 25 |
| 2.2.5. Alternative SNP Ascertainment Methods .....                                                                                   | 26 |
| 2.2.6. STRUCTURE .....                                                                                                               | 26 |
| 2.2.7. Principal Components Analysis (PCA) .....                                                                                     | 27 |
| 2.2.8. Classification by Naïve Bayesian and Multinomial Logistic Regression .....                                                    | 27 |
| 2.3. Results .....                                                                                                                   | 27 |
| 2.3.1. Quality Control .....                                                                                                         | 27 |
| 2.3.2. Genetic Differentiation by $F_{ST}$ .....                                                                                     | 28 |
| 2.3.3. Alternative SNP Ascertainment Methods .....                                                                                   | 29 |
| 2.3.3.1. Method 1: Ranked Mean $F_{ST}$ .....                                                                                        | 30 |
| 2.3.3.2. Method 2: Top 10 and Top 20 Pairwise Comparisons .....                                                                      | 30 |
| 2.3.3.3. Method 3: SNPs with $F_{ST} \geq 0.15$ for 4 Pairwise Comparisons .....                                                     | 31 |
| 2.3.4. STRUCTURE .....                                                                                                               | 31 |
| 2.3.5. Principal Components Analysis (PCA) .....                                                                                     | 32 |
| 2.3.6. BGA Classification .....                                                                                                      | 33 |
| 2.4. Discussion .....                                                                                                                | 34 |

|                                                                                                          |    |
|----------------------------------------------------------------------------------------------------------|----|
| 2.5. References .....                                                                                    | 36 |
| 2.6. Tables .....                                                                                        | 40 |
|                                                                                                          |    |
| CHAPTER 3: Differentiation of Hispanic Biogeographic Ancestry with 80 Ancestry Informative Markers ..... | 42 |
| 3.0. Abstract .....                                                                                      | 43 |
| 3.1. Introduction .....                                                                                  | 44 |
| 3.2. Materials and Methods .....                                                                         | 47 |
| 3.2.1. Genomic Origins and Admixture in Latinos (GOAL) Dataset .....                                     | 47 |
| 3.2.2. 1000 Genomes (1000G) Dataset .....                                                                | 47 |
| 3.2.3. SNP Ascertainment .....                                                                           | 48 |
| 3.2.4. Imputation .....                                                                                  | 49 |
| 3.2.5. STRUCTURE .....                                                                                   | 50 |
| 3.2.6. Principal Components Analysis (PCA) .....                                                         | 51 |
| 3.2.7. Linkage Disequilibrium (LD) Analysis .....                                                        | 52 |
| 3.2.8. Modeling for the Prediction of Unknowns .....                                                     | 52 |
| 3.2.9. Classification of Unknowns .....                                                                  | 52 |
| 3.2.10. Ethical Approval and Informed Consent .....                                                      | 53 |
| 3.3. Results .....                                                                                       | 54 |
| 3.3.1. Setser80 SNP Panel Evaluation .....                                                               | 54 |
| 3.3.2. Classification of Unknowns .....                                                                  | 55 |
| 3.4. Discussion .....                                                                                    | 57 |

|                                                        |         |
|--------------------------------------------------------|---------|
| 3.5. References .....                                  | 60      |
| 3.6. Acknowledgements .....                            | 65      |
| 3.7. Author Contributions .....                        | 66      |
| 3.8. Additional Information .....                      | 66      |
| 3.9. Tables .....                                      | 67      |
| <br>CHAPTER 4: Conclusions and Future Directions ..... | <br>71  |
| 4.1. Summary .....                                     | 72      |
| 4.2. Future Directions .....                           | 73      |
| 4.2.1. Potential Improvements .....                    | 73      |
| 4.2.2. Sample Size .....                               | 73      |
| 4.2.3. Geography .....                                 | 74      |
| 4.3. Applications .....                                | 75      |
| 4.3.1. Forensic Genetics .....                         | 75      |
| 4.3.2. Genetic Genealogy .....                         | 76      |
| 4.3.3. Clinical Genetics .....                         | 77      |
| 4.4. Ethical Considerations .....                      | 78      |
| 4.5. References .....                                  | 79      |
| <br>APPENDIX .....                                     | <br>83  |
| <br>BIBLIOGRAPHY .....                                 | <br>106 |

## LIST OF TABLES

### **CHAPTER 1: *Background and Introduction of Ancestry Informative Markers for Hispanic Populations***

No Tables

### **CHAPTER 2: *SNP Ascertainment for Heterogeneous Hispanic Populations Using the Genomic Origins and Admixture in Latinos Dataset***

Table 1: Proportions of Country Attributable Mean  $F_{ST}$

Table 2: SNPs from Alternative Methods Present in the Setser Panels

Table 3: STRUCTURE Genetic Proportions of Setser Panels

Supplemental Table S2.1:  $F_{ST}$  Statistic for the Setser80

Supplemental Table S2.2: Country Attributable Mean  $F_{ST}$  for the Setser80

Supplemental Table S2.3: Summary of the  $F_{ST}$  Values of the Setser Panel SNPs

Supplemental Table S2.4: Allele Frequencies for the Setser80 on the GOAL Dataset

### **CHAPTER 3: *Differentiation of Hispanic Biogeographic Ancestry with 80 Ancestry Informative Markers***

Table 1: Genetic Proportions from STRUCTURE

Table 2: Naïve Bayesian Classification Accuracy

Table 3: Positive Predictive Values from Naïve Bayes Analysis

Table 4: MLR Classification Accuracy

Supplementary Table S3.1: Country Attributable Mean  $F_{ST}$  Examples

Supplementary Table S3.2: Description of the Setser80 Panel

Supplementary Table S3.3: MLR Confusion Matrix

Supplementary Table S3.4: Naïve Bayes Classification of Panels of 76 SNPs

Supplementary Table S3.5: MLR Classification of Panels of 76 SNPs

## **CHAPTER 4: *Conclusions and Future Directions***

No Tables

## LIST OF FIGURES

### **CHAPTER 1: *Background and Introduction of Ancestry Informative Markers for Hispanic Populations***

Figure 1: Migration Paths Through the Americas

Figure 2: Patterns of Sex-Biased Admixture in Cuba

Figure 3: Percentage of GWAS Studies of Various Populations

Figure 4: Recombination and Linkage Equilibrium

Figure 5: The Relationship Between  $F$  and the Frequency of the Most Frequent Allele ( $M$ )

### **CHAPTER 2: *SNP Ascertainment for Heterogeneous Hispanic Populations using the Genomic Origins and Admixture in Latinos Dataset***

Figure 1: SNP Ascertainment Schematic

Figure 2:  $F_{ST}$  Distributions of SNPs from Setser80, Setser188, and Setser234

Figure 3: STRUCTURE and PCA of Setser Panels

Figure 4: Effect of SNP Panel Size on BGA Classification Accuracy

### **CHAPTER 3: *Differentiation of Hispanic Biogeographic Ancestry with 80 Ancestry Informative Markers***

Figure 1: Comparison to Other Panels

### **CHAPTER 4: *Conclusions and Future Directions***

No Figures

## CHAPTER 1

### *Background and Introduction of Ancestry Informative Markers for Hispanic Populations*

Casandra Hernandez Setser

John V. Planz

Robert C. Barber

Nicole R. Phillips

Ranajit Chakraborty

Deanna S. Cross

Where are you from? This seemingly innocuous question is deceptively complex with varied levels of meaning. One small question can mean “where do you live?”, “where have you lived?”, “where does your family live?”, “where were you born?”, etc.. With modern geographic mobility, it often means “where are your ancestors from?” The answer to this question may also be different, depending on how far back in time you consider. This is especially true of individuals of Hispanic descent, where the mixing of ancestors from different regions of the world was relatively recent.

Colombian Contact began in 1492 with the arrival of Christopher Colombus in the New World. In combination with the beginning of the Trans-Atlantic Slave Trade in 1525<sup>1</sup>, the New World was populated with Europeans, Africans, and Indigenous Peoples of Asian ancestry already living there. It is this intermixing of populations of very different origins and subsequent interbreeding that gives rise to a phenomenon called admixture. The reintroduction of populations that have been isolated from each other over time is responsible for the creation of admixture such as that seen in Hispanic populations. In fact, all people could be considered admixed at some point in history<sup>2,3</sup>. The most extreme case is the ~3% admixture with a separate species, *Homo neanderthalensis*, that remains in modern *Homo sapiens*<sup>4</sup>.

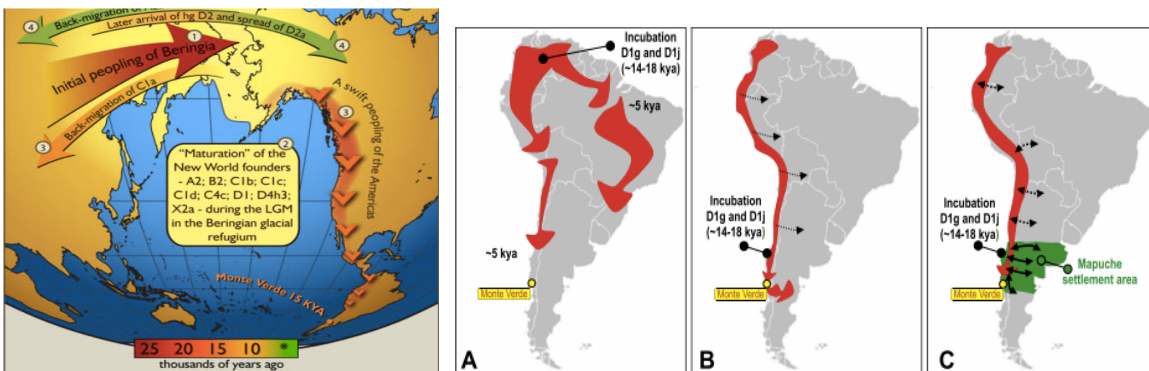
Genetic differentiation is a function of the length of time a population has been isolated and the rate at which mutations accumulate and are passed to the next generation as substitutions. Genetic differentiation between global populations is based on the migration of different groups of people out of Africa and across the world.

## 1.1 Migration and Admixture

### 1.1.2 From East Asians to Indigenous Peoples

For instance, East Asians and Siberians migrated across the Bering Strait at least 15,000 years ago<sup>5-7</sup>, and, with a single nucleotide polymorphism (SNP) mutation rate of  $10^{-8}$  mutations per generation<sup>8</sup>, numerous changes to the genetic code likely occurred during that time.

Subsequently, genetic drift and founder effects differentiated the Asian ancestral populations [often genetically represented by Han Chinese, Beijing (CHB) and Japan, Tokyo (JPT)] from the Indigenous Peoples who settled North and South America. In combination with the 60 mutations per generation *de novo* mutation rate<sup>9</sup> and 28 years per generation<sup>2</sup>, an estimated 32,000 SNPs may have arisen in Native American and Amerindian populations relative to East Asians and Siberians. Since the migration from the Bering Strait to the southern tip of South America took place over time (Figure 1), the SNPs that arose during that time can highlight differences between indigenous populations and, when coupled with their geography, can further refine the likely nationalities of Hispanic individuals by combining genetics with geographic mapping.

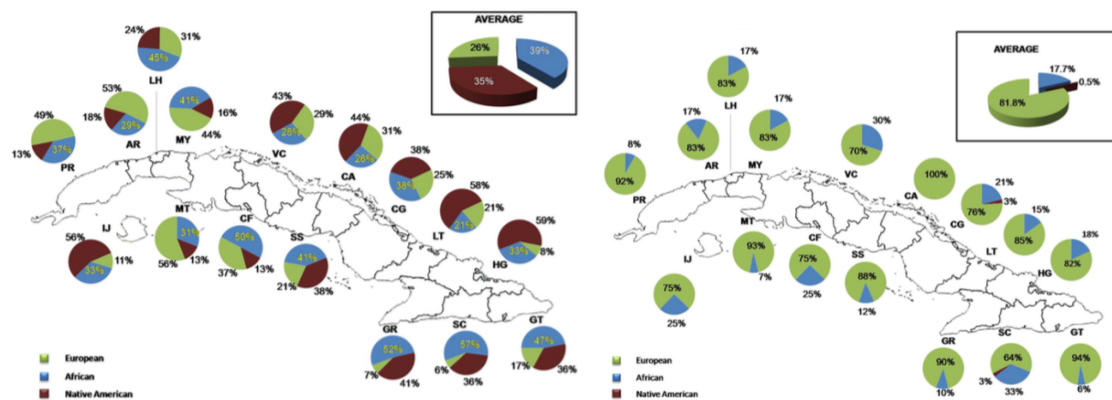


**Figure 1: Migration Paths Through the Americas**

Potential migration routes used as Beringia/Amerindian populations settled in the Americas. This figure depicts evidence from mtDNA which was not tested in my study but still relevant when considering source populations to modern Hispanic populations. Left borrowed from<sup>10</sup> and right borrowed from<sup>11</sup>.

**1.1.3 Hispanic Admixture Complexity**

Hispanic biogeographic ancestry (BGA) is particularly problematic due to the difficulty of parsing out its three-way admixture that often has a limited degree of differentiation between closely related populations<sup>13, 14</sup>. Admixture has traditionally been studied where there are two continental contributing populations, as in African Americans (~70% African Niger-Kordofanian, ~30% European)<sup>15</sup>, which serves as a simpler model for admixture analysis<sup>16, 17</sup>. There are many panels to determine ancestry, though most of the earlier panels are continental in nature, including: Seldin<sup>128</sup><sup>18</sup>, Galanter et al.'s 446<sup>19</sup>, Kidd55<sup>20</sup>, Genetic Atlas<sup>2</sup>, and the Genographic Project<sup>21</sup>. They employ a technique called admixture mapping (MALD = mapping by admixture linkage disequilibrium) whereby they determine percent ancestry for contributing populations (most often by continent) based on origins of segments of chromosomes in linkage disequilibrium<sup>22</sup>. My study focuses on determining biogeographic ancestry (*de facto* country of origin) within the Caribbean, Central and South America directly rather than the percentages of African, Caucasian, and Indigenous contributions as seen in Figure 2.

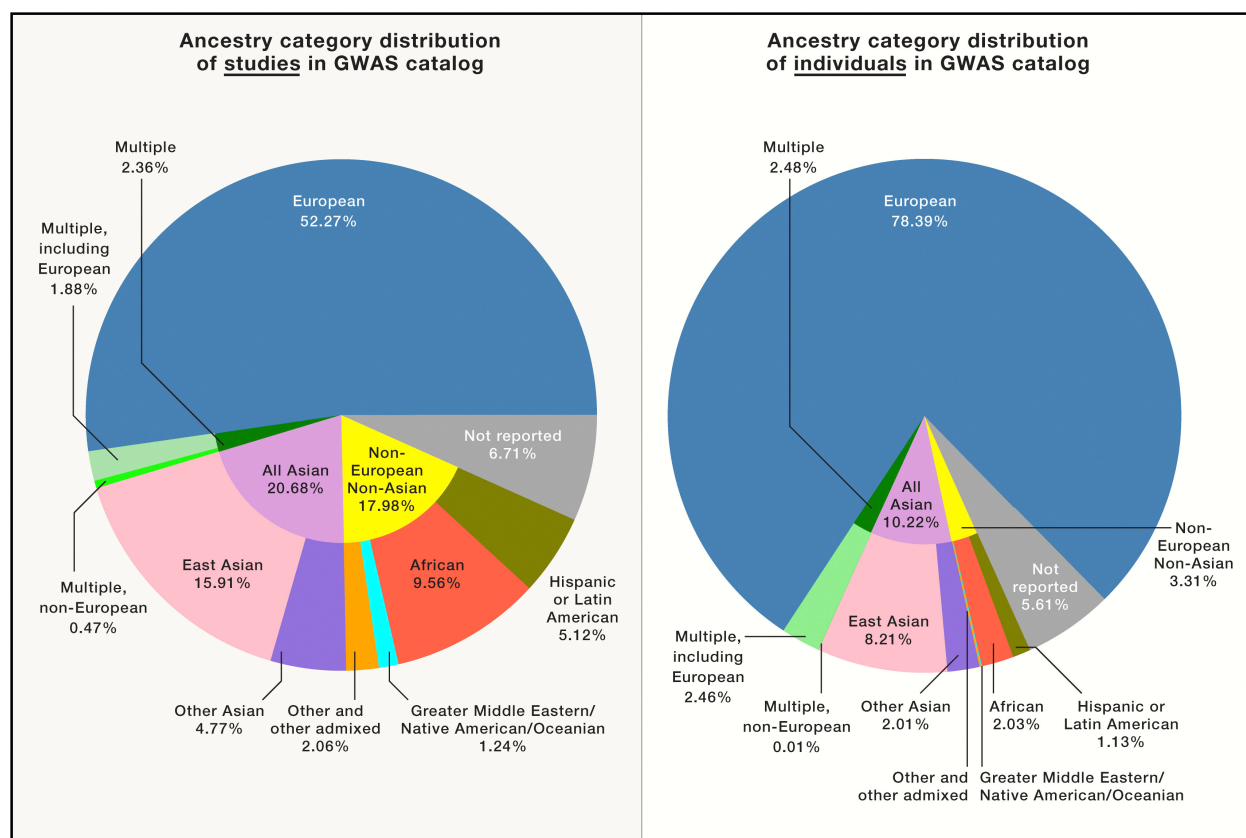


**Figure 2: Patterns of Sex-Biased Admixture in Cuba**

Map of mtDNA contributing population on the left and Y chromosome contributing population on the right. Both images borrowed from<sup>12</sup>.

As the Hispanic population in the United States grows, it has become increasingly important to identify the BGA of heterogeneous Hispanic populations to a similar degree as European populations. European populations have been researched extensively<sup>23</sup> and their relative homogeneity has made these populations easier to study (Figure 3). In the first attempts to capture human genetic variation, HapMap Phase I began with 210 individuals approximately evenly distributed across 3-4 populations: one European population from Utah (CEU), one African population from Ibadan, Nigeria (YRI), and Asian populations represented by Han Chinese, Beijing (CHB) and Tokyo, Japan (JPT)<sup>8</sup>; and no representation of Hispanic populations. The 1000 Genomes Project Phase I also began with 60 of 179 (33.51%) samples from CEU alone (and those may include some of the same individuals)<sup>24</sup>. Of the 2,504 individuals in the current Phase III 1000G super-populations, the European super-population has 503 (20.09%) individuals, while the Admixed American super-population is still under-represented with only 347 (13.86%) individuals<sup>25</sup>. The strategy in documenting human genetic diversity has been to explore the European populations prior to more diverse populations such as Hispanics. Genetic

diversity studies are shifting towards Hispanic populations as more data is available to capture increasingly diverse genetic information and better represent real human populations.



**Figure 3: Percentage of GWAS Studies of Various Populations**

The overwhelmingly lopsided study of Europeans in comparison to any other population. Image borrowed from<sup>23</sup>.

However, SNPs specific to Hispanic populations have not been widely studied, and when they are, they are aimed at a conglomerate population of Hispanics with disparate backgrounds. Lack of appropriate Hispanic reference samples during SNP ascertainment leaves the possibility that loci that are monomorphic in more traditionally studied populations are actually polymorphic in Hispanic populations. Data from Illumina's Infinium Multi-Ethnic AMR/AFR BeadChip<sup>26</sup>, may identify some of these SNPs as it includes populations from Barbados, Brazil,

Colombia, Cuba, Dominican Republic, Honduras, Jamaica, Puerto Rico, etc.<sup>27</sup>. When evaluating differences in populations, treatment plans, etc., more homogeneous samples are easier to separate/distinguish. In general, the effect of variance on statistical significance follows the general biostatistics concept of the bias-variance tradeoff: the higher the variance between compared samples, the lower the probability that those samples will be statistically significant<sup>28</sup>.

## 1.2 Project Design & Chapter Summary

Many of the smaller countries in Mesoamerica and the Caribbean should be relatively homogenous and, aside from large cities, should have limited admixture originating within the last 4-5 generations. I *hypothesize* that it is possible to accurately predict ancestry within heterogeneous Hispanic populations using a small panel of ancestry informative SNPs. Summary level allelic frequencies from this panel can be used to classify specific BGA in the future (see Appendix).

In order to test this, I conducted a series of experiments to determine whether there was sufficient support for my hypothesis. For *Specific Aim 1*, I curated a set of SNPs that can differentiate Hispanic populations. I accomplished this by trimming my dataset for high quality alleles, independent assortment of loci, and genetic differentiation. To increase the efficiency of the panel I selected more SNPs for the low differentiating populations to counteract the ease of separation of the others, as seen in Chapter 2. In *Specific Aim 2*, I compared and contrasted my Hispanic BGA ancestry informative marker (AIMs) panel to pre-existing AIMs panels for efficacy in the differentiation of Hispanic populations, as described in Chapter 3.

I discuss how I created the Setser80 Hispanic AIMs panel in Chapter 2. I used LD and a variant of Wright's  $F_{ST}$ <sup>29</sup> to create a candidate list of 1509 SNPs which showed great genetic differentiation<sup>30</sup> focused on a particular population. Using STRUCTURE<sup>31</sup> and principal

components analysis (PCA), I tested the SNP panel on the SNP ascertainment population, the Genomic Origins and Admixture in Latinos (GOAL)<sup>32</sup>, and 1000 Genomes Admixed American publically available neutral dataset<sup>25</sup>. Based on those results, I refined the panel by allocating more SNPs to the low differentiating populations during panel compression to gain further resolution. I continued forward with the Setser80 as a minimal panel that achieved K = 4 populations in STRUCTURE<sup>31</sup> and retained clustering in PCA.

In Chapter 3, I compared the Setser80, to two highly cited AIMs panels in the literature: the Seldin128<sup>18</sup> and the Kidd55<sup>20</sup> using STRUCTURE analysis, PCA, and micro-simulation. STRUCTURE analysis<sup>31</sup> optimized the Setser80 at K=4, while the Seldin128 and the Kidd55 achieved K=3. Comparing PCA among the three panels, the Setser80 performed better than the Kidd55 which did not show any discernable pattern or clustering. Micro-simulations based on allele frequencies for populations in the GOAL and 1000 Genomes Phase III studies were used to create simulated datasets for all three panels. Some loci from the comparison panels had more than 10% of the individuals missing a genotype and thus that locus was excluded from our micro-simulations, consistent with the QC threshold used in the SNP ascertainment of the Setser80 Hispanic AIMs panel. Using Snipper 2.5 app suite<sup>13</sup> on the micro-simulations, I was able to accurately predict BGA, both by a naïve Bayesian classifier and multinomial logistic regression (MLR) for each panel. The Setser80 demonstrated superior results to the comparison panels for BGA prediction. Additionally, the Setser80 was able to accurately predict BGA on two populations not included in the SNP ascertainment, Peru from Lima (PEL) and Mexicans living in Los Angeles (MXL).

### **1.3 Instruments of Ancestry Determination**

For my analysis of Hispanic BGA, I used various instruments of ancestry determination to select ancestry informative markers. I utilized linkage disequilibrium, which is based on the length of chromosome segments originating from a common ancestor<sup>2</sup>; and  $F_{ST}$ , which is used to determine how different two populations are from each other compared to their internal diversity<sup>31</sup>. I also used two clustering methods: STRUCTURE, which is a commonly used<sup>18, 20, 34, 35</sup> clustering algorithm that decides how many populations are in a dataset and membership of each individual (or portion thereof) in those populations; and PCA which is the gold standard in ancestry analysis<sup>18, 20, 34-37</sup> clustering individuals that are more alike together to explain the greatest amount of variation. Finally, I tested the selected AIMs by Snipper's naïve Bayesian classification and MLR, which are commonly used to predict BGA<sup>13, 38-47</sup>.

#### **1.3.1 Ancestry Informative Markers**

Ancestry informative markers (AIMs) are locations in the DNA that have diverged from each other since the last common ancestor and thus can be used to empirically determine BGA. There are various types of AIMs: insertion/deletions (indels), microhaplotypes, mtDNA, Y chromosome, and SNPs. A SNP is a single point within the DNA sequence where an individual sample may have an alternate nitrogenous base (i.e. allele) than the reference. They arise via random mutations that occur over time, which once passed down through the germ line to the next generation becomes a substitution and contributes to genetic drift of two populations away from each other. Due to their small size, SNPs are particularly useful in the analysis of highly degraded DNA.

The ability of an AIMs panel to differentiate populations relies heavily on the use of relevant reference samples during panel design. The ascertainment process for Hispanic

populations is much more complex when the appropriate reference samples have yet to be created, resulting in the reliance on the continental ancestral lineages (African, European, and Amerindian). The process is complicated due to the three-way admixture which is present in Hispanic individuals. Additionally, Amerindian data is often unavailable and therefore East Asian samples are typically used as proxies<sup>8, 24</sup>. The use of African, European, and East Asian as the three ancestral lineages is problematic, because it does not account for the long period of isolation Indigenous Peoples had from East Asians after the crossing of Beringia<sup>10</sup> or Western Eurasia as a contributor to their genomes<sup>48, 49</sup>.

Previously, researchers focused on global AIMs panels<sup>18-20</sup>, which were calibrated to determine continental origins with a focus across Eurasia, whereas the AIMs described here were designed to supplement global AIMs panels when the sample is presumed to be of Hispanic origin. Elhaik et al. used 40,000 – 130,000 SNPs with complex statistical analysis to differentiate Hispanic ancestries<sup>21</sup>; however, this would be very difficult to implement in most forensic labs or environments with limited resources. In general, there is a trade-off between the greater resolution provided by more SNPs and restricting a panel to a practical size.

Defining continental origin is more straightforward because populations have had thousands of years of genetic history to accumulate substitutions in their DNA, creating alternate alleles with respect to the reference allele. Although continental AIMs have been largely defined in numerous populations, space remains to create additional panels on a finer regional/national scale. Loci that may be monomorphic in one population may in fact have rich polymorphism in another (e.g. rs16891982<sup>41</sup>, rs1800414<sup>41, 50</sup>, rs3811801<sup>50</sup>, and rs671<sup>50</sup>). Additionally, isolated populations may have private alleles that are yet to be identified. Having not found AIMs panels for BGA prediction specific to the Caribbean and Central/South America, I designed my own

panel to predict the specific BGA for Hispanic populations using various bioinformatic tools.

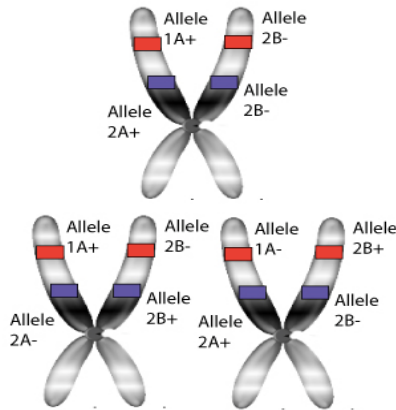
### **1.3.2 Biostatistics & Bioinformatics**

In order to interrogate a dataset for AIMs, an array of tools were used to identify informative SNPs, visualize the capability to separate populations within the dataset, and determine the accuracy of the BGA classification using those SNPs. To ensure the SNPs I selected assorted independently, I filtered the dataset based on LD to create an efficient panel of SNPs that can still distinguish the desired populations. Weir & Cockerham's estimator<sup>51</sup> of Wright's  $F_{ST}$ <sup>29</sup> was calculated as a measure of genetic differentiation in order to select SNPs that were very different between populations. STRUCTURE<sup>31</sup> and PCA are two different clustering algorithms that were used to determine whether the panel had sufficient discrimination of genetic differentiation for the algorithm to separate the populations in the same manner as the actual populations. Finally, we tested the panel's ability to correctly predict Hispanic BGA using a simulated dataset built from the allele frequencies of the actual population data.

#### **1.3.2.1 Linkage Disequilibrium (LD)**

As seen in Figure 4, linkage *equilibrium* is a product of chromosomal recombination during meiosis, where allele 1A+ at locus 1 and 2A- at locus 2 does not predict that the progeny of the parental chromosomes as 1A+, 2A-. The units of linkage are the centimorgan where one centimorgan is the distance between two positions such that 0.01 recombination events occur between them in a single generation. The concrete physical distance corresponding to the centimorgan is 10-160kbp in coding regions, but varies widely across the genome<sup>52</sup>. Linkage disequilibrium (LD) is measured on a scale of 0 to 1 where  $|D'|=1$  represents complete LD<sup>52, 53</sup>. In order to maximize the information contained in a SNP panel, LD is undesirable as the goal is to condense as much information into the fewest number of SNPs<sup>53</sup>, as seen in my Hispanic

AIMs panel. When the SNPs in a panel are in complete linkage equilibrium, and hence segregate independently, the product rule can be applied such that the random match probability is the product of the allele frequencies at each locus.



**Figure 4: Recombination and Linkage Equilibrium**  
Illustration showing how alleles at different loci assort independently due to the process of recombination.  
Borrowed from<sup>54</sup>.

### 1.3.2.2 $F_{ST}$ for Population Differentiation

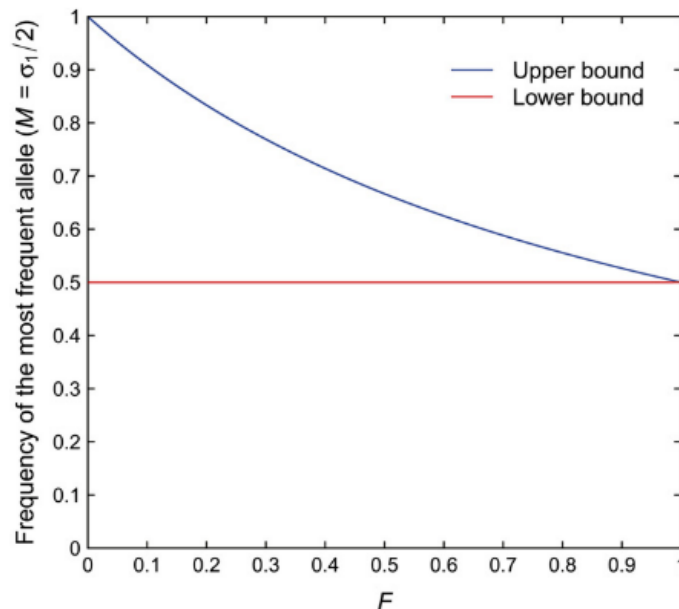
Ding *et al.* compared five measures of genetic differentiation, and reported Informativeness for Assignment Measure  $I_n$  and  $F_{ST}$  statistics performed the best<sup>55</sup>. Conversely, Wright's  $F_{ST}$ <sup>29</sup> is the correlation between gametes within a subpopulation with respect to the total population<sup>33</sup>. Weir & Cockerham extended the  $F_{ST}$  concept to finite populations such as that seen in this study<sup>51</sup>. Allele frequencies are intrinsic to  $F_{ST}$ , specifically the frequency of the most common allele<sup>56</sup> where  $F_{ST}$  is dependent on the diversity of the population. If  $M$  is the allele frequency of the most frequent allele and  $F$  is the genetic differentiation, the highest level of  $F_{ST}$  in the case of two alleles, is where  $M = 0.5$  (Figure 5)<sup>56</sup>. Nei's  $G_{ST}$  (similar to  $F_{ST}$ ) is based on the relationship between the heterozygosity of the total population and the heterozygosity of the sub-population, and complementarily the homozygosity (Eqn 1)<sup>56, 33</sup>. The Wahlund principal

ensures that the homozygosity of the subpopulation will always be greater than that of the total population which, combined with the knowledge that the homozygosity of the sub-population must be less than 1, defines  $G_{ST}$  as greater than 0 and less than 1 (Eqn 2)<sup>33</sup>. Due to the heterozygosity of the total population being less than 1,  $F_{ST}$  is the complement of the ratio of the heterozygosities of the sub-population over the total population and it “cannot exceed the mean homozygosity across subpopulations” (Eqn 3)<sup>33</sup>.

Eqn 1 
$$G_{ST} = \frac{h_T - h_S}{h_T}.$$

Eqn 2 
$$G_{ST} = \frac{H_S - H_T}{1 - H_T}.$$

Eqn 3 
$$F_{ST} = 1 - h_S/h_T < 1 - h_S = H_S.$$



**Figure 5: The Relationship Between  $F$  and the Frequency of the Most Frequent Allele ( $M$ )**

In a two-allele case (binary SNP)  $F_{ST}$  increases as  $M$  approaches 0.5. Borrowed from<sup>56</sup> and reused with permission of the Genetics Society of America.

### **1.3.2.3 Clustering Algorithms**

To visualize the separation of subpopulations, I used STRUCTURE<sup>31</sup> and principal components analysis (PCA) via EIGENSOFT<sup>57</sup>. STRUCTURE<sup>31</sup> is a model-based Bayesian clustering method that estimates the number of populations in a dataset by using an algorithm that includes prior probabilities. PCA is a different clustering method that uses dimensionality reduction to align axes of variation in similar directions into more condensed eigenvectors<sup>58</sup>. Both STRUCTURE<sup>18, 20, 34, 35</sup> and PCA<sup>18, 20, 34-37</sup> are commonly used in the design and evaluation of AIMs panels. I used these two tools to assess each SNP panel's ability to distinguish between the populations provided with the original dataset download.

### **1.3.2.4 Modeling & BGA Classification via Snipper 2.5 App Suite**

Naive Bayesian classification and multinomial logistic regression (MLR)<sup>39</sup> were conducted using the Snipper 2.5 app suite<sup>13</sup>. This web-based classifier was designed for ancestry analysis, particularly in forensics, and allows the user to enter custom spreadsheet training sets<sup>13</sup>, as we have used here. This Bayesian-model based classifier is similar to STRUCTURE<sup>31</sup> but for single samples<sup>13</sup>. While STRUCTURE<sup>31</sup> gives the proportion of membership in each computer determined population, naïve Bayes classifies the sample as belonging to the population with the highest “proportion”. Multiple linear regression uses metric allele frequencies as the independent variable where my desired result is a country name (which is categorical); with logarithmic transformation MLR was a logical choice for BGA prediction<sup>59</sup>. In the past, Snipper 2.5 app suite has most often been used to classify externally visible characteristics [e.g. eye<sup>38, 40, 41</sup> and hair color<sup>42, 43</sup>, and ancestry via the SNPforID 34-plex<sup>44</sup>, Indels<sup>45</sup>, and development<sup>35</sup> and refinement of the EUROFORGEN panel<sup>46, 47</sup>. Ancestry analysis in admixed Hispanic populations has been conducted using Snipper on populations from Bolivia<sup>60, 61</sup>,

Venezuela<sup>40</sup>, as well as Brazil<sup>40</sup>; thus, we considered it robust for classification of admixed populations.

## 1.4 Conclusions

Allele frequencies can be employed to address many different types of problems: forensic, anthropological, pharmacological, clinical, and genealogical. I have created a panel of ancestry informative SNPs that can distinguish BGA down to country sized regions for closely related Hispanic populations. *My research is innovative, because it accurately establishes BGA beyond general ethnicity in a highly efficient SNP panel focused on Hispanic populations that have not been described as extensively as European, African, and Asian populations.*

Although these “races” and/or “ethnicities” are commonly used, scientific enquiry *does not* support the use of externally visible characteristics as a proxy for more subtle ancestry information that is more continuous than discrete<sup>62</sup>. Use of AIMs is more objective and uses algorithms designed to differentiate closely related populations for accurate prediction of BGA. Using two types of clustering software, I designed a SNP panel that accurately predicts BGA of Hispanic populations by two separate classification algorithms. The underlying technique of BGA determination is the allele frequencies of the populations in question. They can be applied in forensics to add weight to the association of crime scene evidence to reference sample and also to determine the origins of anthropological samples and detect possible migration patterns. Allele frequencies can also be used to determine which ancestries are most at risk for which diseases or have deleterious reactions to specific pharmaceuticals. Importantly, this information can be used in clinical studies to ensure that the effects are real and not an artifact of the ancestry of the cases and controls. They are also the foundation of genetic genealogical where they are

used to determine ancestry percentages across the world and potentially find distant relatives.

Although my research began from a forensic point of view, the utility of allele frequency determination applies to a wide variety of genetic questions.

## 1.5 References

- 1) Schroeder, H. *et al.* Genome-wide ancestry of 17th-century enslaved Africans from the Caribbean. *PNAS*. **112** (12), 3669-3673 (2015 March 24). [www.slavevoyages.org](http://www.slavevoyages.org)
- 2) Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science*. **343** (6172), 747-751, (2014 February 14).
- 3) Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*. **513** (7518), 409-413 (2014 September 18).
- 4) Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science*. **343** (6174), 1017-1021 (2014 February 28).
- 5) Rasmussen, M. *et al.* The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature*. **506** (7487), 225-229 (2014 February 13).
- 6) Skoglund, P. *et al.* Genetic evidence for two founding populations of the Americas. *Nature*. **525** (7567), 104-108 (2015 September 3).
- 7) Raghavan, M. *et al.* Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science*. **349** (6250), aab3884; 10.1126/science.aab3884 (2015 August 21).
- 8) The International HapMap Consortium. The International HapMap Project. *Nature*. **426** (6968), 789-796 (2003 December 18).
- 9) Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature*. **488** (7412), 471-475 (2012 August 23).
- 10) Tamm, E., *et al.* Beringian standstill and spread of Native American founders. *PLoS ONE*. **2** (9), e829; 10.1371/journal.pone.0000829 (2007 September 5).
- 11) Bodner, M. *et al.* Rapid coastal spread of First Americans: Novel insights from South America's Southern Cone mitochondrial genomes. *Genome Res*. **22** (5), 811-820 (2012 May).

- 12) Marcheco-Teruel, B. *et al.* Cuba: Exploring the history of admixture and genetic basis of pigmentation using autosomal and uniparental markers. *PLoS Genet.* **10** (7), e1004488; 10.1371/journal.pgen.1004488 (2014 July 14).
- 13) Phillips, C. *et al.* Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int-Gen.* **1** (3-4), 273–280 (2007 December).
- 14) Price, A. L. *et al.* A genomewide admixture map for Latino populations. *Am J Hum Genet.* **80** (6), 1024-1036 (2007 June).
- 15) Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science.* **324** (5930), 1035-1044 (2009 May 22).
- 16) Gurdasani, D. *et al.* The African genome variation project shapes medical genetics in Africa. *Nature.* **517** (7534), 327-332 (2015 January 15).
- 17) Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *Am J Hum Genet.* **96** (1), 37–53 (2015 January 8).
- 18) Kosoy, R. *et al.* Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat.* **30** (1), 69-78 (2009 January).
- 19) Galanter, J. M. *et al.* Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genet.* **8** (3), e1002554; 10.1371/journal.pgen.1002554 (2012 March).
- 20) Kidd, K. K. *et al.* Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci Int-Gen.* **10** (1), 23-32 (2014 May).
- 21) Elhaik, E. *et al.* Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat Commun.* **5**, 3513; 10.1038/ncomms4513 (2014 April 29).
- 22) Patterson, N. *et al.* Methods for high-density admixture mapping of disease genes. *Am J Hum Genet.* **74** (5), 979-1000 (2004 May).
- 23) Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell.* **177** (1), 26-31 (2019 March 21).
- 24) 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature.* **467** (7319), 1061-1073 (2010 October 28).
- 25) Auton, A. *et al.* A global reference for human genetic variation. *Nature.* **526** (7571), 68–74 (2015 October 1).

- 26) Infinium Multi-Ethnic AMR/AFR BeadChip data sheet. Illumina. Pub. No. 370-2015-006. (2016 February 29).
- 27) Johnston, H. R. *et al.* Identifying tagging SNPs for African specific genetic variation from the African Diaspora Genome. *Sci Rep.* **7**, 46398; 10.1038/srep46398 (2017 April 21).
- 28) Singh, S. Understanding the bias-variance tradeoff. (2018 May 20). URL <https://towardsdatascience.com/understanding>
- 29) Wright, S. The genetical structure of populations. *Ann Eugen.* **15** (4), 323-354 (1951).
- 30) Hartl, D. L. & Clark, A. G. (3<sup>rd</sup> ed.) Principles of Population Genetics. (Sinauer Associates, Inc., 1997).
- 31) Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics.* **155** (2), 945-959 (2000 June).
- 32) Moreno-Estrada, A. *et al.* Reconstructing the population genetic history of the Caribbean. *PLoS Genet.* **9** (11), e1003925; 10.1371/journal.pgen.1003925 (2013 November 14).
- 33) Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating, and interpreting  $F_{ST}$ . *Nat Rev Genet.* **10** (9), 639-650 (2009 September).
- 34) Kidd, J., Friedlaender, F. R., Speed, W. C., Pakstis, A. J., De La Vega, F. M. & Kidd, K. K. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investig Genet.* **2** (1), 1; 10.1186/2041-2223-2-1 (2011 January 5).
- 35) Phillips, C. *et al.* Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set. *Forensic Sci Int-Gen.* **11** (1), 13–25 (2014 July).
- 36) Paschou, P., Lewis, J., Javed, A. & Drineas, P. Ancestry informative markers for fine-scale individual assignment to worldwide populations. *J Med Genet.* **47** (12), 835-847 (2010 December).
- 37) Huerta-Chagoya, A. *et al.* A panel of 32 AIMs suitable for population stratification correction and global ancestry estimation in Mexican mestizos. *BMC Genetics.* **20** (1), 5; 10.1186/s12863-018-0707-7 (2019 January 8).
- 38) Ruiz, Y. *et al.* Further development of forensic eye color predictive tests. *Forensic Sci Int-Gen.* **7** (1), 28-40 (2013 January).

- 39) McNevin, D. *et al.* An assessment of Bayesian and multinomial logistic regression classification systems to analyse admixed individuals. *Forensic Sci Int-Gen. Supplement Series* **4** (1), e63-e64; 10.1016/j.fsigss.2013.10.032 (2013).
- 40) Freire-Aradas, A. *et al.* Exploring iris colour prediction and ancestry inference in admixed populations of South America. *Forensic Sci Int-Gen.* **13**, 3-9 (2014 November).
- 41) Maroñas, O. *et al.* Development of a forensic skin colour predictive test. *Forensic Sci Int-Gen.* **13**, 34-44 (2014 November).
- 42) Söchtig, J. *et al.* Exploration of SNP variants affecting hair colour prediction in Europeans. *Int J Legal Med.* **129** (5), 963–975 (2015 September).
- 43) Pośpiech, E. *et al.* The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction. *Forensic Sci Int-Gen.* **11**, 64-72 (2014 July).
- 44) Fondevila, M. *et al.* Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies. *Forensic Sci Int-Gen.* **7** (1), 63-74 (2013 January).
- 45) Pereira, R. *et al.* Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. *PLoS ONE.* **7** (1), e29684; 10.1371/journal.pone.0029684 (2012 January 17).
- 46) De la Puente, M. *et al.* The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs. *Forensic Sci Int-Gen.* **22**, 81–88 (2016 May).
- 47) Eduardoff, M. *et al.* Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM™. *Forensic Sci Int-Gen.* **23**, 178-189 (2016 July 1).
- 48) Reich, D. *et al.* Reconstructing Native American population history. *Nature.* **488** (7411), 370-374 (2012 August 16).
- 49) Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature.* **505** (7481), 87-91 (2014 January 2).
- 50) Santangelo, R., González-Andrade, F., Børstinga, C., Torroni, A., Pereira, V. & Morling, N. Analysis of ancestry informative markers in three main ethnic groups from Ecuador supports a trihybrid origin of Ecuadorians. *Forensic Sci Int-Gen.* **31**, 29-33 (2017 November).
- 51) Cockerham, C. C. & Weir, B. S. Estimation of gene flow from F-Statistics. *Evolution.* **47** (3), 855-863 (1993 June).

- 52) Reich, D. *et al.* Linkage disequilibrium in the human genome. *Nature*. **411** (6834), 199-204 (2001 May 10).
- 53) Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L. & and Nickerson, D. A. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet*. **74** (1), 106–120 (2004 January).
- 54) Khan, R. Basic concepts - linkage disequilibrium. *Gene Expression*. (2007 January 24). URL <https://www.gnXP.com/WordPress/2007/01/24/basic-concepts-linkage-disequilibrium/>
- 55) Ding, L. *et al.* Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC Genomics*. **12**, 622; 10.1186/1471-2164-12-622 (2011 December 20).
- 56) Jakobsson, M., Edge, M. D. & Rosenberg, N. A. The relationship between  $F_{ST}$  and the frequency of the most frequent allele. *Genetics*. **193** (2), 515-528 (2013 February).
- 57) Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet*. **2** (12), 2074–2093; 10.1371/journal.pgen.0020190 (2006 December 22).
- 58) Jeong, D. H., Ziemkiewicz, C., Ribarsky, W. & Chang, R. Understanding principal component analysis using a visual analytics tool. Charlotte Visualization Center. (UNC Charlotte, 2008).
- 59) McDonald, J. H. (3<sup>rd</sup> ed.) Multiple Regression, 229-237. Handbook of Biological Statistics. (Sparky House Publishing, 2014). URL <http://www.biostatHandbook.com/multipleregression.html>
- 60) Heinz, T. *et al.* Ancestry analysis reveals a predominant Native American component with moderate European admixture in Bolivians. *Forensic Sci Int-Gen*. **7** (5), 537–542 (2013).
- 61) Taboada-Echalar, P. *et al.* The genetic legacy of the pre-colonial period in contemporary Bolivians. *PLoS ONE*. **8** (3), e58980; 10.1371/journal.pone.0058980 (2013 March).
- 62) Yudell, M., Roberts, D., DeSalle, R. & Tishkoff, S. Taking race out of human genetics. *Science*. **351** (6273), 564-565 (2016 February 5).

## CHAPTER 2

### *SNP Ascertainment for Heterogeneous Hispanic Populations using the Genomic Origins and Admixture in Latinos Dataset*

Cassandra Hernandez Setser

John V. Planz

Robert C. Barber

Nicole R. Phillips

Ranajit Chakraborty

Deanna S. Cross

## 2.1 Introduction

The overall goal of this project was to design a panel of ancestry informative markers (AIMs) that could separate heterogeneous Hispanic populations. A panel to differentiate country-sized populations into their respective biogeographic ancestries (BGA) has not been accomplished for Hispanic populations. Using data from five populations of the Genomic Origins and Admixture in Latinos (GOAL)<sup>1</sup> study, I designed a set of AIMs panels that were able to correctly predict BGA with limited information about the individual. I hypothesized that it was possible to accurately assign specific BGA within Hispanic populations using a small set of ancestry informative markers.

There are nearly as many methods for single nucleotide polymorphism (SNP) ascertainment as there are published AIMs panels. Older panels use large numbers of SNPs (e.g. Price et al., 2007)<sup>2</sup>, while modern panels use high quality SNPs that are carefully selected by various methods. Candidate SNPs may be selected using raw allele frequency differential across populations  $\delta$  (delta)<sup>3, 4</sup>,  $F_{ST}$ <sup>5-7</sup>, Rosenberg's  $I_n$ <sup>6, 7</sup>, or SNP weights from principal components analysis (PCA)<sup>7, 8</sup>. Their candidate SNPs have been narrowed down further using locus specific branch length (LSBL)<sup>5</sup> or receiver operating curves (ROC)<sup>7</sup>.

While older ancestry estimations rely on a large number of SNPs to accurately predict BGA<sup>2, 9, 10</sup>, more recently, smaller curated SNP panels have demonstrated their utility. As more information on global genetic diversity has become available {e.g. Human Genome Diversity Project – Centre d'Etude du Polymorphisme Humain (HGDP-CEPH)<sup>11</sup>, HapMap<sup>12</sup>, and The 1000 Genomes Project<sup>13</sup>}, research groups have carefully selected, condensed SNP panels and compared them to panels using large numbers (~100,000) of randomly selected SNPs to demonstrate that high quality SNPs provide similar results<sup>14, 15</sup>. C. Phillips et al. designed a

panel of 34 AIMs by selecting SNPs with substantial allele frequency differences between three populations<sup>3</sup>. Paschou et al. demonstrated that, using a hierarchical decision tree based approach, they could achieve comparable results with as few as 0.1% of the original 650,000 SNPs<sup>16</sup>. Galanter et al. employed locus-specific branch length (LSBL) using  $F_{ST}$  statistics and neighbor joining trees to estimate genetic distance<sup>5, 17</sup>. They selected 446 AIMs for admixture mapping of Latin American populations and distinguished populations within 10% of the values from genome-wide data, for 95% of the samples<sup>5</sup>. Huerta-Chagoya et al. used the first principal component SNP weights to differentiate European and Native American ancestral contributions within Mexican mestizos and found their population discernment with 32 AIMs<sup>8</sup> comparable to previously existing panels<sup>5, 14, 18</sup>.

These methods all yielded viable SNP panels. For comparison, Zeng et al. tested three statistical methods ( $\delta$ ,  $F_{ST}$ , and  $I_n$ ) to design AIMs panels using Ancestry SNPMiner<sup>19</sup> in combination with PCA and receiver operating curves (ROC)<sup>7</sup>. Utilizing the HapMap III data<sup>20</sup> on African Americans, Caucasians, East Asians, and Hispanic Americans; they performed PCA on all permutations of the four populations and three statistical methods to select the top 30 SNPs each. The quality of clustering for each panel was determined using PCA plots and ROC curves to select diagnostic/prediction thresholds that maximize sensitivity (true positive rate) while minimizing the complement of specificity (false positive rate) and define the PC 1 cutoff value<sup>21</sup>.

Once researchers established the accuracy of BGA with smaller, high quality panels, the field expanded to admixture proportions<sup>22</sup> and admixture mapping<sup>2</sup>. Such panels could differentiate populations, depending on the SNP selection method. Various AIMs panels have been published on specific regions of the world including Brazil<sup>23</sup>, Pacifiplex<sup>24</sup>, Ecuador<sup>25</sup>, Malaysia<sup>26</sup>, etc. Given the increasing number of panels, C. Phillips gives a comprehensive

overview of AIMs selection methods<sup>27</sup>. In 2016, Soundararajan et al. reviewed 21 existing ancestry informative SNP panels and found only 46 of the 1397 total SNPs were present in three or more panels<sup>28</sup>.

My Hispanic AIMs panel uses a variation of Wright's  $F_{ST}$  method<sup>29</sup>, which focuses on the four pairwise comparisons with a country in common (country attributable mean  $F_{ST}$ ), for a more granular BGA when Hispanic ancestry is presumed. I chose  $F_{ST}$  statistics based on the results from Ding et al. due to 1) better correlation than other AIMs statistics<sup>6</sup>, 2) direct incorporation of the degree of heterozygosity<sup>30</sup>, and 3) it is more common in recent literature<sup>5, 7, 31</sup>. For my panels, all SNPs where country attributable mean  $F_{ST} \geq 0.15$ , were ranked by highest mean  $F_{ST}$  and balanced based on the number of SNPs attributed to each country as 1<sup>st</sup> or 2<sup>nd</sup> country attributable mean  $F_{ST}$ .

## **2.2 Materials and Methods**

### **2.2.1 The Genomic Origins and Admixture in Latinos (GOAL) Study**

I used 160 unrelated individuals from the GOAL study whose samples were previously collected in South Florida as family trios in which three of four grandparents were from the same country<sup>1</sup>. My chosen populations consisted of Honduras (HUR,  $n = 13$ ), the Dominican Republic (DOM,  $n = 21$ ), Colombia (COL,  $n = 53$ ), Cuba (CUB,  $n = 55$ ), and Puerto Rico (PUR,  $n = 18$ )<sup>1</sup> and includes SNP data from 897,336 autosomal SNPs across all chromosomes with paired phenotypic and genotypic information.

### **2.2.2 SNP Ascertainment**

SNPs were filtered for linkage disequilibrium (LD), missingness, and minor allele frequency (maf) using PLINK v.1.9<sup>32</sup>. To maximize the value of the entire panel, SNPs were

removed if  $LD \geq 0.7$  and did not assort independently of other SNPs, leaving 531,878 SNPs. Additionally, I removed SNPs where the number of samples not genotyped at that locus was  $\geq 0.10$  so that BGA determination was not made on incomplete information for the remaining 522,083 SNPs. Further quality control included filtering out SNPs with a minor allele frequency ( $maf$ )  $< 0.01$ , resulting in a dataset of 494,886.

### **2.2.3 Overview of $F_{ST}$**

$F_{ST}$  statistics were calculated for the ten pairwise comparisons at each of 494,886 SNPs and 1,578 SNPs were identified where at least one pairwise  $F_{ST} \geq 0.14$ . The 1,509 SNPs that exceeded  $F_{ST} \geq 0.15$  for at least one pairwise comparisons had their  $F_{ST}$  statistics averaged by country (e.g. HUR where HUR vs. DOM, HUR vs. COL, HUR vs. CUB, and HUR vs. PUR were averaged) which became the country attributable mean  $F_{ST}$  for that country. The country with the highest country attributable mean  $F_{ST}$  per SNP was assigned as the 1<sup>st</sup> country and the next highest as the 2<sup>nd</sup> country.

### **2.2.4 AIMs Panel Creation**

From the 1,509, I selected 234 SNPs using the country attributable mean  $F_{ST}$  and pared the panel down to 80 SNPs. The Setser80 excludes those SNPs with paired HUR and DOM as 1<sup>st</sup> and 2<sup>nd</sup> country. Additionally, I removed SNPs with lower 1<sup>st</sup> country attributable mean  $F_{ST}$  SNPs for HUR and DOM and retained those SNPs where COL or CUB was the 1<sup>st</sup> country.

Finally, I verified that the AIMs I selected were not present in other AIMs panels to ensure their utility as a supplementary panel. I compared my AIMs to those used in multiple studies<sup>14, 28, 31</sup>, both directly and by  $LD \leq 0.7$  using the LDMatrix function within the LDLink website<sup>33</sup> hosted by National Cancer Institute.

### **2.2.5 Alternative SNP Ascertainment Methods**

Although I devised country attributable mean  $F_{ST}$  as a way to broaden the search for potentially informative SNPs per country, there are other common methods used in population genetics studies. First, I calculated the  $F_{ST}$  values of ~500,000 SNPs and ranked them. For method 1, I selected the top 234 SNPs from the 1509 based solely on mean  $F_{ST}$ , or the “Top 234 mean  $F_{ST}$ ”. In method 2, I selected the Top 20 and Top 10 pairwise  $F_{ST}$  values for each pairwise comparison; there were 99 unique SNPs in the “99 Top 20 SNPs and 51 unique SNPs in the “51 Top 10 SNPs”. For method 3, I selected any SNP with  $F_{ST} \geq 0.15$  for any pairwise comparison, recalculated the  $F_{ST}$ , and chose those that had any four or five pairwise comparisons (not exclusively country attributable mean  $F_{ST}$ ) above threshold as the “131 SNPs from 4 Pairwise”.

### **2.2.6 STRUCTURE**

Using a Bayesian model-based clustering algorithm (STRUCTURE) v. 2.3.4<sup>34</sup>, I assessed whether the SNP panels could distinguish populations. In this program, each vertical line represents an individual and the different colors represent the computationally determined ancestry proportions for each individual. STRUCTURE<sup>34</sup> analysis was conducted using 10,000 burn-in and 100,000 Monte Carlo Markov Chain (MCMC) repetitions with 10 iterations per level of K (K=2 to K=7) for the two Setser panels<sup>34</sup>. I applied the Evanno method<sup>35</sup> via STRUCTURE Harvester<sup>36</sup> at each of the computer’s estimated number of populations (K), and selected those with the most likely number of clusters (populations). The final STRUCTURE diagrams for each SNP panel were aligned by CLUMPP<sup>37</sup> and averaged by Distruct<sup>38</sup>.

Quantifiable genetic proportions, which underlie the STRUCTURE<sup>34</sup> algorithm, provide continuous, quantifiable genetic proportions using the pre-defined K level to determine the number of clusters. For the most likely level of K<sup>35</sup>, I aligned the populations and averaged the

ten iterations in order to report ancestry proportion clusters per population. The underlying numerical genetic proportions served to quantify the degree of genetic difference of each population.

### **2.2.7 Principal Components Analysis (PCA)**

Each dataset was also evaluated by PCA, a dimensionality reduction clustering method, using EIGENSOFT v.6.1.4<sup>39</sup> and plotted using the first three eigenvectors. Genesis<sup>40</sup> was used for improved visualization of clustering.

### **2.2.8 Classification by Naïve Bayesian and Multinomial Logistic Regression**

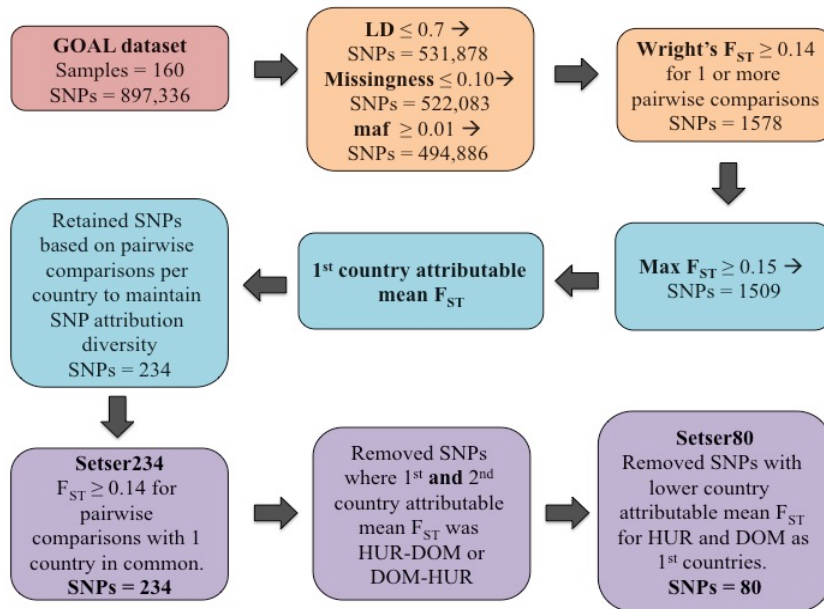
For the prediction of biogeographic ancestry, I utilized the web-based program Snipper 2.5 app suite<sup>3</sup>. This program uses  $-\log(\text{likelihood})$  as the basis for a naïve Bayesian classifier and multinomial logistic regression.

## **2.3 Results**

### **2.3.1 Quality Control**

I used PLINK v.1.9<sup>32</sup> to prune the dataset of  $n=244$  individuals, retaining  $n=160$  unrelated individuals and unlinked, high quality SNPs with great genetic differentiation<sup>41</sup>. The GOAL study was collected as family trios; therefore I removed 84 “children” to maximize the size of an unrelated dataset to  $n=160$ <sup>1</sup>. Using Plink v.1.9<sup>32</sup>, I filtered for  $LD \leq 0.7$  and reduced the 897,336 autosomal SNPs to 531,878. For quality control, I filtered out missingness  $\geq 0.1$ , yielding 522,083 SNPs, and  $maf < 0.01$ , yielding 494,886 autosomal SNPs. Using Weir & Cockerham’s estimator<sup>42</sup>, I selected the 1509 SNPs where  $F_{ST} \geq 0.15$  for at least one pairwise comparison. Of the 1509, I selected 234 SNPs and a subset of 80 SNPs (Figure 1).

**Figure 1: SNP Ascertainment Schematic**



**Figure 1: SNP Ascertainment Schematic**

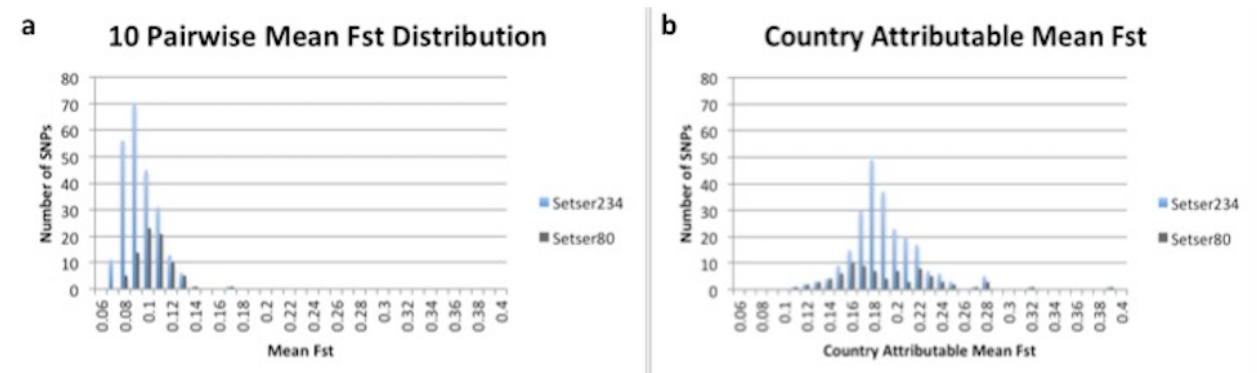
Methods used to select the SNPs used in the Setser234 and Setser80. Abbreviations used: SNPs = single nucleotide polymorphisms, LD = linkage disequilibrium, maf = minor allele frequency, Max = maximum, HUR = Honduras, DOM = Dominican Republic, COL = Colombia, CUB = Cuba, and PUR = Puerto Rico.

### 2.3.2 Genetic Differentiation by $F_{ST}$

The  $F_{ST}$  distributions of the Setser234 and Setser80 were indistinguishable by mean  $F_{ST}$  across all ten pairwise comparisons (Figure 2a) and country attributable mean  $F_{ST}$  (Figure 2b). The mean of the mean  $F_{ST}$  values for the Setser234 was (mean 0.09441, StDev = 0.01461) and was (mean = 0.10516, StDev = 0.01469) for the Setser80, with rs12431505 mean  $F_{ST}$  = 0.17 (Figure 2a)(Supplemental Table S2.1). Comparing the country attributable mean  $F_{ST}$ , there was no appreciable difference in mean as the panel was condensed from Setser234 (country attributable mean  $F_{ST}$  = 0.19124, StDev = 0.03228) to Setser80 (country attributable mean  $F_{ST}$  = 0.19174, StDev = 0.04754) (Figure 2b). The two outliers, rs12435621 and rs12431505, were

members of both panels with country attributable mean  $F_{ST} = 0.31$  and  $0.39$ , respectively (Figure 2b)(Supplemental Table S2.2).

**Figure 2:  $F_{ST}$  Distributions of SNPs from Setser234 and Setser80**



**Figure 2:  $F_{ST}$  Distributions of SNPs from Setser234 and Setser80**

a) Distribution of SNPs by mean  $F_{ST}$  across all ten pairwise comparisons for both panels.  
b) Distribution of SNPs by country attributable mean  $F_{ST}$  across four pairwise comparisons of both panels. Abbreviations used: SNP = single nucleotide polymorphism.

The proportion of overall SNPs with  $F_{ST} \geq 0.15$  was stable at 32.6% to 33.3%, indicating the Setser SNP panels retained high quality SNPs during panel compression (Supplemental Table S2.3). Additionally, high differentiating populations HUR (35% to 23.1%) and DOM (22.6% to 13.8%) had reduced proportions of attributable SNPs from Setser234 to Setser80 while COL (13.5% to 25%) and CUB (16% to 27.5%) increased (Table 1).

### 2.3.3 Alternative SNP Ascertainment Methods

To further verify my SNP panels, I selected SNPs by three alternative methods: by highest mean  $F_{ST}$ , by the top 10 – 20  $F_{ST}$  values for each pairwise comparison, and the other selecting any SNP with  $F_{ST} \geq 0.15$  for any pairwise comparison and a subset where any four pairwise comparisons had  $F_{ST} \geq 0.15$ .

### **2.3.3.1 Method 1: Ranked Mean $F_{ST}$**

For comparison, I selected the top 234 SNPs from the 1509 solely by highest mean  $F_{ST}$  and 145 of these SNPs were present in the Setser234 (Table 2). Of the SNPs present only in the Setser234, an additional 21 were attributed to PUR demonstrating that Setser234 was able to select SNPs to distinguish populations with lower genetic differentiation. The 234 SNPs with the highest mean  $F_{ST}$  (mean  $F_{ST}$  = 0.09 - 0.17) would have excluded SNPs like rs3910480 where mean  $F_{ST}$  = 0.08, but was attributed to COL (1<sup>st</sup> country attributable mean  $F_{ST}$  = 0.14) and CUB (2<sup>nd</sup> country attributable mean  $F_{ST}$  = 0.07), two populations requiring additional SNPs to resolve.

### **2.3.3.2 Method 2: Top 10 and Top 20 Pairwise Comparisons**

Starting with the dataset of 494,886 SNPs, I selected the SNPs with the 20 highest  $F_{ST}$  values for each pairwise comparison. After removing duplicates, there were 99 Top 20 SNPs, 82 of which were present in the Setser234. For the Top 20 SNPs, the SNPs present in the Setser234 comprised 35% of the panel and in the Setser80 comprised 67.5% of the panel (Table 2). Of the 99 Top 20 SNPs, the 17 not captured by Setser234 were identified from the following comparisons: HUR vs. COL = 3, HUR vs. PUR = 1, COL vs. CUB = 4, CUB vs. PUR = 9, and 1 shared by HUR vs. COL and COL vs. CUB. Three of these did not have one pairwise  $F_{ST} \geq 0.15$  and were not present in the 1509. The others did not meet my criteria because they either did not have enough pairwise comparisons above threshold, the pairwise comparisons did not have a country in common, or would have been filtered out in the Setser80 as a lower value HUR attributed SNP. Similarly, there were 51 Top 10 SNPs, 49 of which were present in the Setser234 comprising 20.9% of the Setser234 and 43.8% of the Setser80 (Table 2).

### 2.3.3.3 Method 3: SNPs with $F_{ST} \geq 0.15$ for 4 Pairwise Comparisons

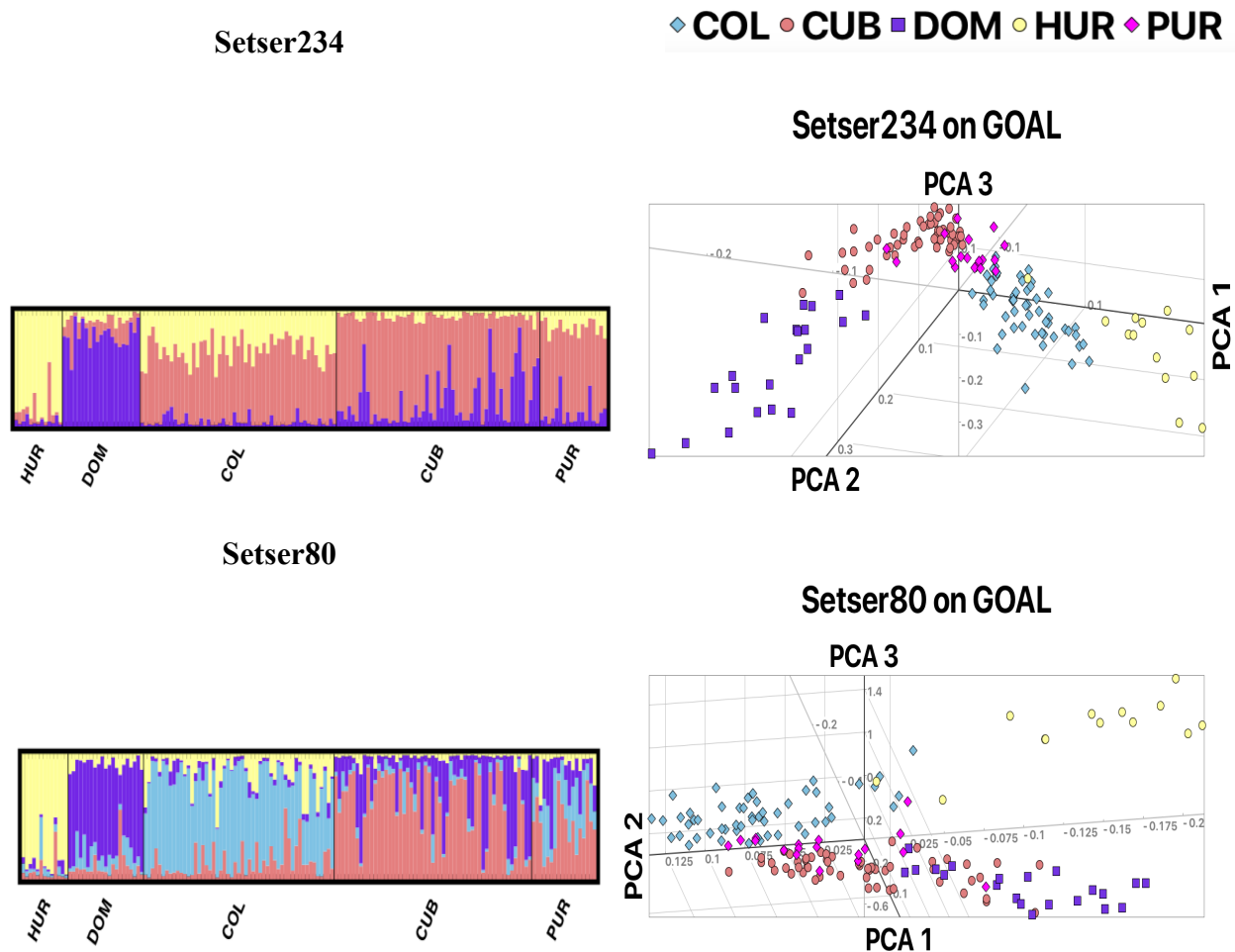
I selected SNPs with  $F_{ST} \geq 0.15$  for any four pairwise comparisons with no regard to country resulting in 131 SNPs that met criteria, of which 129 were already in the Setser234. These 131 SNPs covered 55.1% of the Setser234 and 62.5% of the Setser80 SNPs, which indicated the panel was streamlined from the Setser234 to the Setser80. Using the country attributable  $F_{ST}$  method, a large proportion of the SNPs identified by the three alternative methods were still captured.

### 2.3.4 STRUCTURE

To determine the optimum level of K and utility for population differentiation, I interrogated the Setser234 using a Bayesian model-based clustering method, STRUCTURE<sup>34</sup>, followed by the Evanno method<sup>35</sup> (as applied in STRUCTURE Harvester<sup>36</sup>) (Figure 3a). The Setser234 at K=3 identified three distinct clusters with different predominant genetic proportions: HUR (Cluster 1 = 0.8529), DOM (Cluster 2 = 0.7888), and COL|CUB|PUR (Table 3). It was difficult to determine whether PUR clustered more closely with COL or CUB since all three had their highest proportion in Cluster 3: COL (Cluster 3 = 0.6199, Cluster 1 = 0.3376), PUR (Cluster 3 = 0.7378, Cluster 1 = 0.1465), and CUB (Cluster 3 = 0.7203, Cluster 2 = 0.235) (see Table 3).

As determined by the Evanno method<sup>35</sup>, the Setser80 optimized to K=4 where COL was differentiated as Cluster 3, (distinguishing it from CUB|PUR) to reveal the final four clusters: HUR (Cluster 1 = 0.8290), DOM (Cluster 2 = 0.6976), COL (Cluster 3 = 0.6562), and CUB|PUR (Cluster 2 = 0.2892|0.2048, Cluster 3 = 0.0634|0.2969, and Cluster 4 = 0.6125|0.4145) (Table 3).

**Figure 3: STRUCTURE and PCA of Setser Panels**



**Figure 3: STRUCTURE and PCA of Setser Panels**

Each plot represents 160 unrelated GOAL individuals and their respective populations. Figures a and c are STRUCTURE<sup>34</sup> plots where each vertical line represents one person. Figures b and d are PCA plots created through EIGENSOFT<sup>39</sup> where the first three principal components are plotted. Figures a and b use the Setser234 SNP panel (K=3) while c and d use the Setser80 (K=4). Abbreviations used: HUR = Honduras, DOM = Dominican Republic, COL = Colombia, CUB = Cuba, PUR = Puerto Rico, and PCA = principal components analysis.

### 2.3.5 Principal Components Analysis (PCA)

In the PCA of Setser234, HUR, DOM, COL, and CUB|PUR were easily identifiable (Figure 3b). Therefore, separating these three regions using PCA was straightforward: Central America (HUR), South America (COL), and Caribbean (DOM, CUB and PUR). The Dominican Republic (DOM) clustered better in STRUCTURE<sup>34</sup> as Cluster 2, (Figure 3a), but overlapped

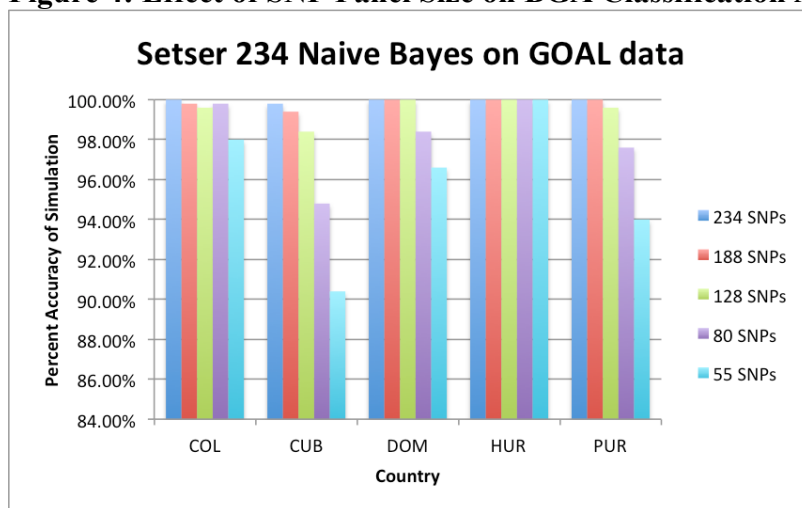
more with CUB in PCA as I condensed the panel (Figure 3d). In contrast, COL clustered better in PCA (Figure 3b), but only officially became Cluster 3 in the STRUCTURE<sup>34</sup> analysis of Setser80, K=4 (Figure 3d). The decreased clustering of DOM and increased clustering of COL in PCA (Figure 3d) as the panel was compressed from 234 to 80 SNPs is a relic of removing SNPs attributed to HUR and DOM while enriching for low differentiating populations.

By observing the first three principal components, it is evident that the Setser80 quantitatively provides clusters that are less ambiguous than the Setser234. The increased separation is most clear in the first principal component (PC1) where Setser80 PC1=27.8% and Setser234 PC1=22.4%.

### **2.3.6 BGA Classification**

Utilizing the allele frequencies calculated above, I created five sets of 500 micro-simulations via a resampling approach to model the data<sup>43</sup>. I evaluated the validity the Setser panels using the naïve Bayesian classifier and multinomial logistic regression (MLR) classifier found in Snipper 2.5 app suite<sup>3</sup> to predict Hispanic BGA. With the nested SNP panels, I used Snipper to test the accuracy of prediction of each of country in each panel size. Additionally, I tested an intermediate panel of 188 SNPs as well as panels of 128 and 55 SNPs to mimic the size of the 55 SNPs from Kidd et al., 2014 and the 128 SNPs from Kosoy et al., 2009 for comparison purposes (see Chapter 3)<sup>14, 31</sup>. I found that HUR and COL were robust at as few as 55 SNPs with an average 98% accuracy (Figure 4). For the remaining three countries, there was greater than 95% accuracy using 80 SNPs. While 128 SNPs would have resulted in 2-3% greater accuracy, the Setser80 balanced small panel size (>100 SNPs) with 95% accuracy.

**Figure 4: Effect of SNP Panel Size on BGA Classification Accuracy**



**Figure 4: Effect of SNP Panel Size on BGA Classification Accuracy**

Percent accuracy of naïve Bayes classification in the Snipper 2.5 app suite<sup>3</sup> for five different sized panels in each of five countries. Snipper selected SNPs for each subset based on degree of divergence. Abbreviations used: COL = Colombia, CUB = Cuba, DOM = Dominican Republic, HUR = Honduras, PUR = Puerto Rico, SNP = single nucleotide polymorphism.

## 2.4 Discussion

I designed panels of 234 and 80 AIMs capable of differentiating Hispanic individuals using a variant of Wright's  $F_{ST}$ <sup>29</sup> based on five populations captured in the Genomic Origins and Admixture in Latinos (GOAL) study<sup>1</sup>. This was accomplished using country attributable mean  $F_{ST}$ , which averages the four comparisons with one country in common rather than the ten possible pairwise comparisons. To test my method for choosing SNPs, I compared my AIMs panel to other methods employed by various research groups<sup>44,45</sup> and most of the SNPs from these methods were selected in my panel(s).

Comparing the two AIMs panels, the Setser80 performs comparably (or better) to the Setser234 with fewer SNPs. The  $F_{ST}$  distributions revealed no appreciable difference in mean  $F_{ST}$  between the two panels (Figure 2). The PCA results exhibited similar clustering and the concentration of genetic differentiation into fewer SNPs and resulted in higher eigenvalues for

the first PCs (Figure 3). The 68.5% reduction in panel size to increase the contribution of low differentiating populations resulted in the formation of a fourth cluster in the STRUCTURE<sup>34</sup> diagram (Figure 3). Furthermore, the genetic proportions behind the STRUCTURE algorithm<sup>34</sup> provided additional support regarding the retention of genetic differentiation during panel compression (Table 3).

My Setser80 panel may be an addition to existing panels. The EUROFORGEN Global AIM-SNP panel from C. Phillip et al., which leveraged data mining by reviewing ten separate studies to select 128 SNPs<sup>15</sup>, did not encompass any of the Setser SNPs. Shortly thereafter, Soundararajan et al. conducted an expanded review of 21 published AIMs panels for overlap; not a single one of the Setser SNPs were included in the resulting consensus panel<sup>28</sup>. Independent assortment of the Setser80 SNPs from those identified by Soundararajan et al.<sup>28</sup> and C. Phillips et al.<sup>15</sup> fulfills the goal of supplementing global panels; where no adjustments for redundancy are necessary. Therefore, the Setser80 may be used as a whole, without removal of AIMs in LD with the AIMs from the primary panel.

Although the Setser AIMs panels reliably classified BGA, this study has its limitations: sample size vs. relatedness, accurate representation of BGA, the inaccuracy of self-identification of ancestry, and utilization of an early generation chip<sup>46</sup> for genotyping. This study used data from the GOAL study developed by Moreno-Estrada et al. which is available for download from the Database of Genotypes and Phenotypes (dbGAP)<sup>1</sup>. Its 250 samples (6 Haitians removed for small sample size) were collected as family trios in South Florida. Therefore, I removed ~1/3 of the individuals due to relatedness to prevent false enrichment of certain alleles. Also, data from the GOAL study was a convenience sample where: 1) an immigrant population may not be truly representative of their source BGA and 2) self-identification of ancestry can often be inaccurate.

While this study provides new and useful information on the complex genetics of heterogeneous Hispanic populations, it also creates more avenues of enquiry. The logical next step would be to test the portability of this panel into another (larger) dataset, such as the American Admixed super-population from 1000 Genomes (n = 347)<sup>47</sup>. In the future, the expansion of my study would use data from an updated genechip (e.g. Infinium Multi-Ethnic AMR/AFR beadchip from Illumina)<sup>48</sup> or obtain sequence data from similar populations. Sequence data would be invaluable to the identification of additional loci that are polymorphic in Hispanic populations, but monomorphic in more extensively studied (e.g. European) populations. Importantly, the degree of characterization of a population factors into the level of influence it has on the design of the next generation of gene-chips and AIMs panels.

## 2.5 References

- 1) Moreno-Estrada, A. *et al.* Reconstructing the population genetic history of the Caribbean. *PLoS Genet.* **9** (11), e1003925; 10.1371/journal.pgen.1003925 (2013 November 14).
- 2) Price, A. *et al.* A genomewide admixture map for Latino populations. *Am J Hum Genet.* **80** (6), 1024-1036 (2007 June).
- 3) Phillips, C. *et al.* & The SNPforID Consortium. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int-Gen.* **1** (3-4), 273-280 (2007 December).
- 4) Wathen, M. J., Gautam, Y, Ghandikota, S., Rao, M. B. & Marsha, T. B. LEI: A novel allele frequency-based Feature Selection method for multi-ancestry admixed populations. *Sci Rep.* **9** (1), e11103; 10. 1038/s41598-019-47012-y (2019 July 31).
- 5) Galanter, J. M. *et al.* Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genet.* **8** (3), e1002554; 10.1371/journal.pgen.1002554 (2012 March).
- 6) Ding, L. *et al.* Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC Genomics.* **12**, 622; 10.1186/1471-2164-12-622 (2011 December 20).

- 7) Zeng, X., Chakraborty, R., King, J. L., LaRue, B., Moura-Neto, R. S. & Budowle, B. Selection of highly informative SNP markers for population affiliation of major US populations. *Int J Legal Med.* **130** (2), 341–352 (2016 March).
- 8) Huerta-Chagoya, A. *et al.* A panel of 32 AIMs suitable for population stratification correction and global ancestry estimation in Mexican mestizos. *BMC Genetics.* **20** (1), 5; doi.org/10.1186/s12863-018-0707-7 (2019 January 8).
- 9) Tian, C. *et al.* Analysis of East Asia genetic substructure using genome-wide SNP arrays. *PLoS ONE.* **3** (12), e3862; 10.1371/journal.pone.0003862 (2008 December 5).
- 10) Elhaik, E. *et al.* Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat Commun.* **5**, 3513; 10.1038/ncomms4513 (2014 April 29).
- 11) Cann, H. M. *et al.* A human genome diversity cell line panel. *Science.* **296** (5566), 261-262 (2002 April 12).
- 12) The International HapMap Consortium. The International HapMap Project. *Nature.* **426** (6968), 789-796 (2003 December 18).
- 13) The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature.* **467** (7319), 1061-1073 (2010 October 28).
- 14) Kosoy, R. *et al.* Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat.* **30** (1), 69-78 (2009 January).
- 15) Phillips, C. *et al.* Building a forensic ancestry panel from the ground up: the EUROFORGEN Global AIM-SNP set. *Forensic Sci Int-Gen.* **11** (1), 13-25 (2014 July).
- 16) Paschou, P., Lewis, J., Javed, A. & Drineas, P. Ancestry informative markers for fine-scale individual assignment to worldwide populations. *J Med Genet.* **47** (12), 835-847; (2010 December).
- 17) Shriver, M. *et al.* The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum Genomics.* **1** (4), 274-286 (2004 May).
- 18) Lai, C.-Q. *et al.* Population admixture associated with disease prevalence in the Boston Puerto Rican health study. *Hum Genet.* **125** (2), 199-209 (2009).
- 19) Amirisetty, S., Hershey, G. K. K. & Baye, T. M. AncestrySNPminer: A bioinformatics tool to retrieve and develop ancestry informative SNP panels. *Genomics.* **100** (1), 57-63 (2012 July).

- 20) Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature*. **467** (7311), 52-58 (2010 September 2).
- 21) Zeng, X. Selection of highly informative markers for apportionment of ancestry and population affiliation. Fort Worth, TX: University of North Texas Health Science Center. (2016 May 1).
- 22) Moreno-Estrada, A. *et al.* The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science*. **344** (6189), 1280-1285 (2014 June 13).
- 23) Santos, H. C. *et al.* A minimum set of ancestry informative markers for determining admixture proportions in a mixed American population: the Brazilian set. *Eur J Hum Genet*. **24** (5), 725-731 (2016 May 1).
- 24) Santos, C. *et al.* Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific Region. *Forensic Sci Int-Gen*. **20**, 71-80 (2016 January 1).
- 25) Santangelo, R., Gonzalez-Andrade, F., Børsting, C., Torroni, A., Pereira, V. & Morling, N. Analysis of ancestry informative markers in three main ethnic groups from Ecuador supports a trihybrid origin of Ecuadorians. *Forensic Sci Int-Gen*. **31**, 29-33 (2017 November).
- 26) Yahya, P. *et al.* Analysis of the genetic structure of the Malay population: Ancestry-informative marker SNPs in the Malay of Peninsular Malaysia. *Forensic Sci Int-Gen*. **30**, 152–159 (2017 September).
- 27) Phillips C. Forensic genetic analysis of biogeographic ancestry. *Forensic Sci Int-Gen*. **18**, 49-65 (2015).
- 28) Soundararajan, U., Yun, L., Shi, M. & Kidd, K. K. Minimal SNP overlap among multiple panels of ancestry informative markers argues for more international collaboration. *Forensic Sci Int-Gen*. **23**, 25-32 (2016 July 1).
- 29) Wright, S. The genetical structure of populations. *Ann Eugen*. **15** (4), 323-354 (1951 March).
- 30) Jakobsson, M., Edge, M. D. & Rosenberg, N. A. The relationship between  $F_{ST}$  and the frequency of the most frequent allele. *Genetics*. **193** (2), 515-528 (2013 February).
- 31) Kidd, K. K. *et al.* Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci Int-Gen*. **10** (1), 23-32 (2014 May).
- 32) Purcell, S. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. **81** (3), 559-575 (2007 September).

- 33) Machiela, M. J. & Chanock, S. J. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*. **31** (21), 3555-3557 (2015 December 18). URL <https://ldlink.nci.nih.gov>
- 34) Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics*. **155** (2), 945-959 (2000 June).
- 35) Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol Ecol*. **14** (8), 2611-2620 (2005 July).
- 36) Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour*. **4** (2), 359-361 (2012).
- 37) Jakobsson, M. & Rosenberg, N. A. CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*. **23** (14), 1801-1806 (2007 July 15).
- 38) Rosenberg, N. A. Distruct: A program for the graphical display of population structure. *Mol Ecol Notes*. **4** (1), 137-138 (2004 March).
- 39) Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet*. **2** (12), 2074-2093; 10.1371/journal.pgen.0020190 (2006 December).
- 40) Buchmann, R.W. Genesis: Copyright © 2014, University of the Witwatersrand.
- 41) Hartl, D. L. & Clark, A. G. (3<sup>rd</sup> ed.) Principles of Population Genetics. (Sinauer Associates, Inc., 1997).
- 42) Cockerham, C. C. & Weir, B. S. Estimation of gene flow from F-Statistics. *Evolution*. **47** (3), 855-863 (1993 June).
- 43) Yuan, X., Miller, D. J., Zhang, J., Herrington, D. & Wang, Y. An overview of population genetic data simulation. *J Comput Biol*. **19** (1), 42-54 (2012 January 1).
- 44) Nievergelt, C. M. *et al.* Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investig Genet*. **4** (1), Article 13 (2013 July 13).
- 45) Das, R. & Upadhyai, P. An ancestry informative marker set which recapitulates the known fine structure of populations in South Asia. *Genome Biol Evol*. **10** (9), 2408-2416 (2018 September 1).

- 46) McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet.* **40** (10), 1166-1174 (2008 October).
- 47) Auton, A. *et al.* A global reference for human genetic variation. *Nature.* **526** (7571), 68-74 (2015 September 30).
- 48) Infinium Multi-Ethnic AMR/AFR BeadChip data sheet. Illumina. Pub. No. 370-2015-006. (2016 February 29).

## 2.6 Tables

**Table 1: Proportions of Country Attributable Mean  $F_{ST}$**

|                                                                        | Setser234 % (n) | Setser80 % (n) |
|------------------------------------------------------------------------|-----------------|----------------|
| <b>HUR 1<sup>st</sup> + 2<sup>nd</sup> Country Proportion (# SNPs)</b> | 35% (164)       | 23.1% (37)     |
| <b>DOM 1<sup>st</sup> + 2<sup>nd</sup> Country Proportion (# SNPs)</b> | 22.6% (106)     | 13.8% (22)     |
| <b>COL 1<sup>st</sup> + 2<sup>nd</sup> Country Proportion (# SNPs)</b> | 13.5% (63)      | 25% (40)       |
| <b>CUB 1<sup>st</sup> + 2<sup>nd</sup> Country Proportion (# SNPs)</b> | 16% (75)        | 27.5% (44)     |
| <b>PUR 1<sup>st</sup> + 2<sup>nd</sup> Country Proportion (# SNPs)</b> | 12.8% (60)      | 10.6% (17)     |

**Table 1: Proportions of Country Attributable Mean  $F_{ST}$**

Description of the two Setser panels by the number of SNPs attributed to each country by 1<sup>st</sup> and 2<sup>nd</sup> highest country attributable mean  $F_{ST}$ . The proportion of SNPs attributed to each country per panel is reported with the number of SNPs given in parentheses (). Note that each SNP is counted twice: once for 1<sup>st</sup> country attribution and once for 2<sup>nd</sup> country attribution.

**Table 2: SNPs from Alternative Methods Present in the Setser Panels**

|           | Method 1                       | Method 2                |                         | Method 3                          |
|-----------|--------------------------------|-------------------------|-------------------------|-----------------------------------|
| Panel     | Top 234<br>Mean $F_{ST}$ % (n) | 99 Top 20<br>SNPs % (n) | 51 Top 10<br>SNPs % (n) | 131 SNPs from<br>4 Pairwise % (n) |
| Setser234 | 62% (145)                      | 35% (82)                | 20.9% (49)              | 55.1% (129)                       |
| Setser80  | 91.3% (73)                     | 67.5% (54)              | 43.8% (35)              | 62.5% (50)                        |

**Table 2: SNPs from Alternative Methods Present in the Setser Panels** Number of SNPs present in the Setser234 and Setser80 that were selected by the three alternative methods. The proportion of SNPs selected by each alternative method that is present in the Setser234 and the Setser80; number of SNPs is listed in parentheses (). Note that the Top 20 SNPs are in addition to those already identified in Top 10. Proportions are listed in relation to the number of SNPs in the Setser panel (e.g.  $50/80 = 62.5\%$ ).

**Table 3: STRUCTURE Genetic Proportions for Setser Panels on the GOAL Dataset**

| Panel           | Population | Cluster 1     | Cluster 2     | Cluster 3     | Cluster 4     | n  |
|-----------------|------------|---------------|---------------|---------------|---------------|----|
| Setser80 (K=4)  | HUR        | <b>0.8290</b> | 0.0387        | 0.0647        | 0.0676        | 13 |
| Setser80 (K=4)  | DOM        | 0.0811        | <b>0.6976</b> | 0.1147        | 0.1067        | 21 |
| Setser80 (K=4)  | COL        | 0.1601        | 0.0474        | <b>0.6562</b> | 0.1365        | 53 |
| Setser80 (K=4)  | CUB        | 0.0348        | 0.2892        | 0.0634        | <b>0.6125</b> | 55 |
| Setser80 (K=4)  | PUR        | 0.0836        | 0.2048        | 0.2969        | <b>0.4145</b> | 18 |
| Setser234 (K=3) | HUR        | <b>0.8529</b> | 0.0496        | 0.0976        | N/A           | 13 |
| Setser234 (K=3) | DOM        | 0.0717        | <b>0.7888</b> | 0.1396        | N/A           | 21 |
| Setser234 (K=3) | COL        | 0.3376        | 0.0424        | <b>0.6199</b> | N/A           | 53 |
| Setser234 (K=3) | CUB        | 0.0448        | 0.2350        | <b>0.7203</b> | N/A           | 55 |
| Setser234 (K=3) | PUR        | 0.1465        | 0.1158        | <b>0.7378</b> | N/A           | 18 |

**Table 3: STRUCTURE Genetic Proportions for Setser Panels on the GOAL Dataset** Proportion of each population's separation into the computer calculated clusters as determined by the Evanno method (K=3 or K=4). Clusters were aligned per population and averaged across the ten iterations. Abbreviations used: HUR = Honduras, DOM = Dominican Republic, COL = Colombia, CUB = Cuba, and PUR = Puerto Rico.

## CHAPTER 3

### *Differentiation of Hispanic Biogeographic Ancestry with 80 Ancestry Informative Markers*

Published in Scientific Reports  
May 2020

Cassandra H. Setser  
John V. Planz  
Robert C. Barber  
Nicole R. Phillips  
Ranajit Chakraborty  
Deanna S. Cross

### 3.0 Abstract

Ancestry informative single nucleotide polymorphisms (SNPs) can identify biogeographic ancestry (BGA); however, population substructure and relatively recent admixture can make differentiation difficult in heterogeneous Hispanic populations. Utilizing unrelated individuals from the Genomic Origins and Admixture in Latinos dataset (GOAL, n=160), we designed an 80 SNP panel (Setser80) that accurately depicts BGA through STRUCTURE and PCA. We compared our Setser80 to the Seldin and Kidd panels via resampling simulations, which models data based on allele frequencies. We incorporated Admixed American 1000 Genomes populations (1000G, n=347), into a combined populations dataset to determine robustness. Using multinomial logistic regression (MLR), we compared the 3 panels on the combined dataset and found overall MLR classification accuracies: 93.2% Setser80, 87.9% Seldin panel, 71.4% Kidd panel. Naïve Bayesian classification had similar results on the combined dataset: 91.5% Setser80, 84.7% Seldin panel, 71.1% Kidd panel. Although Peru and Mexico were absent from panel design, we achieved high classification accuracy on the combined populations for Peru (MLR = 100%, naïve Bayes = 98%), and Mexico (MLR = 90%, naïve Bayes = 83.4%) as evidence of the portability of the Setser80. Our results indicate the Setser80 SNP panel can reliably classify BGA for individuals of presumed Hispanic origin.

### 3.1 Introduction

It is important to study the genetics of Hispanic populations to avoid oversimplifying this heterogeneous ethnicity into a single conglomerate. The identification of specific biogeographic ancestries (BGA) has implications both in clinical<sup>1</sup> and forensic<sup>2</sup> genetics. Clinically, a more complete description of the various Hispanic BGAs may result in identification of rare variants that may not have been previously described when grouping all Hispanic populations together<sup>3</sup>, or for controlling for population substructure in clinical trials<sup>4, 5</sup>. Hispanic individuals are known to have differential predispositions for various diseases and ignoring this diversity restricts the generalizability of the results<sup>6</sup>. In forensics, BGA data could be used to investigate the origin of unidentified human remains (UHR)<sup>7</sup>, or locate the rightful parents/guardians of a child who is unable to identify where she/he is from<sup>8</sup>. It is the heterogeneous nature of Hispanic populations that has previously deterred full characterization of their substructure. However, in the past decade, there has been a movement to explore global human diversity and a variety of genetic panels have been designed for this purpose.

Early ancestry informative marker (AIMs) panels are “continental” in nature, focused on admixture mapping to determine from which of the six inhabited continents an individual has ancestry; these include: Seldin128<sup>9</sup>, Galanter et al.’s 446<sup>10</sup>, Kidd55<sup>11</sup>, EUROFORGEN<sup>12</sup>, Genetic Atlas<sup>13</sup>, Genographic Project<sup>14</sup>, Cuba by Marcheco-Teruel et al.<sup>15</sup>, and Cuba by Fortes-Lima et al.<sup>16</sup>. Although these studies assessed continental ancestry proportions (e.g. Seldin128)<sup>9</sup>, highly differentiated populations may be detected within continental panels, even identifying admixed populations such as Gujarati Indians in Houston, TX and Mexican ancestry from Los Angeles, CA<sup>17</sup>. The ability to separate small admixed populations among larger more homogenous populations supports the notion that continental SNPs with high genetic differentiation may still

be informative on a more specific country level. The simultaneous description of highly divergent populations alongside less specific populations using the same SNP panel is central to the goals of our study. However, dual level analysis of admixed populations within continental panels is rare, as it tends to decrease the panel's performance<sup>2, 17</sup>.

Other panels target more specific, country BGA beginning in European populations before extending to other regions of the world (e.g. Denmark within Northern Europe). Although the Genographic Project<sup>14</sup> assessed populations worldwide (though sparsely in the Americas), their in-house geographic population structure (GPS) algorithm is capable of identifying country of origin. EASTASAIMS was one of the first non-European AIMS panels focusing on 22 East Asian populations using 1,500 AIMS and was able to separate the five largest populations in the region<sup>18</sup>. Zeng et al.<sup>19</sup> created a panel of 23 AIMS using  $F_{ST}$  focusing on the four major US populations from HapMap 3<sup>20</sup>: African ancestry from Southwest United States (ASW), Utah residents with Northern and Western European ancestry (CEU), Chinese from Metropolitan Denver, Colorado (CHD), and Mexican ancestry from Los Angeles, CA (MEX). And more recently, Huerta-Chagoya et al.<sup>21</sup> reported 32 AIMS within Mexican mestizo populations, to estimate admixture proportions in various regions of Mexico.

Highly accurate BGA predictions are possible with up to 83% accuracy, but at the expense of panel size, requiring 40,000 – 130,000 SNPs as used in the Genographic Project<sup>14</sup>. Additionally, of the 12,476 reference samples used to select 40,000+ SNPs in their panel, only 9% were from American/Amerindian populations<sup>22</sup>, which limits the utility of their panel for resolving Hispanic ancestry. The size of this panel<sup>14</sup>, the proprietary nature of the SNPs on their Genochip<sup>22</sup>, and poor representation of the Western hemisphere, has prompted us to create a

small, efficient, and publicly available SNP panel concentrated on BGA of Central America, South America, and the Caribbean.

Within one country, both Great Britain<sup>23</sup> and Cuba<sup>15, 16</sup> have attempted to describe the diversity of their populations. The British Isles were ideal candidates for national differentiation due to their relative homogeneity and the presence of a geographic barrier which has historically restricted continuous gene flow with continental Europe and other island populations. In contrast, studies by Marcheco-Teruel et al.<sup>15</sup>, and Fortes-Lima et al.<sup>16</sup> superficially appear to differentiate between the fifteen Cuban provinces on a national level, but their real focus was measuring admixture proportions using a subset of Galanter et al.'s 446 SNPs<sup>10</sup>, making their studies better described as continental and highlighting the need for a within country panel. Overall, at least 21 AIMs panels have been reported; however, of the 1,397 SNPs identified by Soundararajan et al.<sup>24</sup>, only 46 Consensus SNPs were in common to three or more SNP panels.

At present, there is no AIMs panel that focuses on the determination of BGA between countries in the Americas. Despite the overlap of our region of interest with the Galanter et al.'s 446 Latin American AIMs<sup>10</sup>, our purpose was to classify BGA, not to estimate the ancestral proportions contributed from 3-4 continental populations. The majority of AIMs panels and genetic ancestry studies have a heavy concentration of populations in Europe and Asia and far fewer in Central America, South America, and the Caribbean<sup>13, 14, 18</sup>. Our country panel addresses this gap in knowledge and focuses on these same populations.

## **3.2 Materials and Methods**

### **3.2.1 Genomic Origins and Admixture in Latinos (GOAL) Dataset**

Here we downloaded the GOAL dataset and used 160 unrelated individuals including Honduran (HUR, n=13), Dominican Republican (DOM, n=21), Colombian (COL, n=53), Cuban (CUB, n=55), and Puerto Rican (PUR, n=18) populations with three of four grandparents from the same country<sup>25</sup>. These samples were collected in South Florida and genotyped using the Affymetrix 6.0 gene chip of 906,600 predetermined SNPs<sup>26</sup>.

The Genomic Origins and Admixture in Latinos (GOAL) dataset analyzed during the current study is available in the dbGaP repository, accession number phs000750.v1.p1, found at: [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000750.v1.p1&phv=202273&phd=4443&pha=&pht=3936&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000750.v1.p1&phv=202273&phd=4443&pha=&pht=3936&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1). Funding support for the GOAL Study was provided by the National Institute of General Medical Sciences (1R01GM090087). Additional support for sample collection was provided by a grant from the Stanley J. Glaser Foundation and the Dr. John T. Macdonald Foundation Department of Human Genetics.

### **3.2.2 1000 Genomes (1000G) Dataset**

For further comparison, we used fully sequenced individuals from the 1000 Genomes Project Phase 3 Admixed American populations (n=347)<sup>27</sup>, accessed through the UCSC Genome Browser<sup>28</sup>. These include Colombia in Medellin (CLM, n=94), Peru in Lima (PEL, n=85), Puerto Rico (PUR, n=104), and Mexican Living in Los Angeles (MXL, n=64)<sup>27</sup>. The 1000 Genomes Project dataset is available via the UCSC Genome Browser<sup>28</sup>, found at: <http://genome.ucsc.edu/>.

### 3.2.3 SNP Ascertainment

We created our AIMs panel by applying a series of quality control algorithms. Beginning with 897,336 autosomal SNPs on the genechip<sup>26</sup>, we filtered the GOAL dataset by linkage disequilibrium ( $LD \leq 0.7$ , missingness  $\leq 0.1$ , and minor allele frequency ( $maf \geq 0.01$ ) using PLINK v.1.9<sup>29-30</sup> and retained 494,886 SNPs. After calculating  $F_{ST}$ <sup>31</sup> by Weir & Cockerham's algorithm<sup>32</sup> in PLINK v.1.9 ([https://www.cog-genomics.org/plink/1.9/basic\\_stats#fst](https://www.cog-genomics.org/plink/1.9/basic_stats#fst))<sup>33</sup>, 1509 SNPs with  $F_{ST} \geq 0.15$  for at least one pairwise comparison were retained.

We calculated the mean  $F_{ST}$  for each of the five countries and assigned each SNP to a country based on the highest mean  $F_{ST}$ . The next highest mean  $F_{ST}$  was designated the 2<sup>nd</sup> country mean  $F_{ST}$ . For example, rs3777908 is attributed to HUR because the average of HUR vs. DOM, HUR vs. COL, HUR vs. CUB, and HUR vs. PUR is  $[(0.27318 + 0.19754 + 0.19560 + 0.28808)/4] = 0.23860$ , which was the highest mean  $F_{ST}$  value for rs3777908. The 2<sup>nd</sup> highest mean  $F_{ST} = 0.07442$ , corresponded to PUR (see Supplemental Table S3.1 for example calculations).

We binned the 1509 SNPs by the 1<sup>st</sup> and 2<sup>nd</sup> highest country attributable mean  $F_{ST}$  and removed SNPs where the 1<sup>st</sup> country mean  $F_{ST} < 0.11$  and 2<sup>nd</sup> country mean  $F_{ST} < 0.09$ , resulting in 437 SNPs. Since 63.3% of the 1509 candidate SNPs were attributable to HUR or DOM, we removed SNPs where HUR and DOM had the 1<sup>st</sup> and 2<sup>nd</sup> highest country mean  $F_{ST}$ , where HUR had the 2<sup>nd</sup> highest country mean  $F_{ST}$ , and the 100 lowest ranked SNPs where HUR or DOM had the highest country mean  $F_{ST}$ . From the remaining 247 SNPs, we chose a subset of 80 in order to maintain ~20% contribution of SNPs for each country across 1<sup>st</sup> and 2<sup>nd</sup> country attribution.

Therefore we proceeded with the Setser80 (Supplemental Table S3.2), which has the following country attributable mean  $F_{ST}$  values: HUR (mean  $F_{ST} = 0.21228$ ), DOM (mean  $F_{ST} =$

0.16901), COL (mean  $F_{ST}$  = 0.14212), CUB (mean  $F_{ST}$  = 0.10803), and PUR (mean  $F_{ST}$  = 0.10272).

To assess the value of our panel, we compared it to two commonly sited AIMs panels<sup>9, 11</sup>. Here, we refer to the panel developed by Kosoy et al., 2009 as the Seldin128<sup>9</sup>, and the 55 ancestry informative SNPs developed by Kidd et al., 2014 as the Kidd55<sup>11</sup>. We performed each analysis on the Setser80 in parallel with the Kidd and Seldin panels to evaluate the utility of our Hispanic AIMs panel.

### **3.2.4 Imputation**

The SNPs on the Affymetrix 6.0 gene chip<sup>26</sup> were pre-determined and not all SNPs were included in the ABI Taqman assay used to genotype the Seldin128<sup>9</sup> and Kidd55<sup>11</sup>; therefore, we imputed these two panels into the GOAL dataset<sup>25</sup> using IMPUTE2<sup>34</sup> on the full 250 individuals using a 5Mb window centered on each SNP and an effective population size of 20,000 as seen in Instructions for IMPUTE version 2<sup>35</sup>. We used 2,504 individuals from 1000G<sup>27</sup> for the genetic map and legend and the strand alignment from dbSNP batch query. Given the use of genome builds NCBI35/hg17 to GRCh38/hg38, we converted all components to GRCh37/hg19 for analysis.

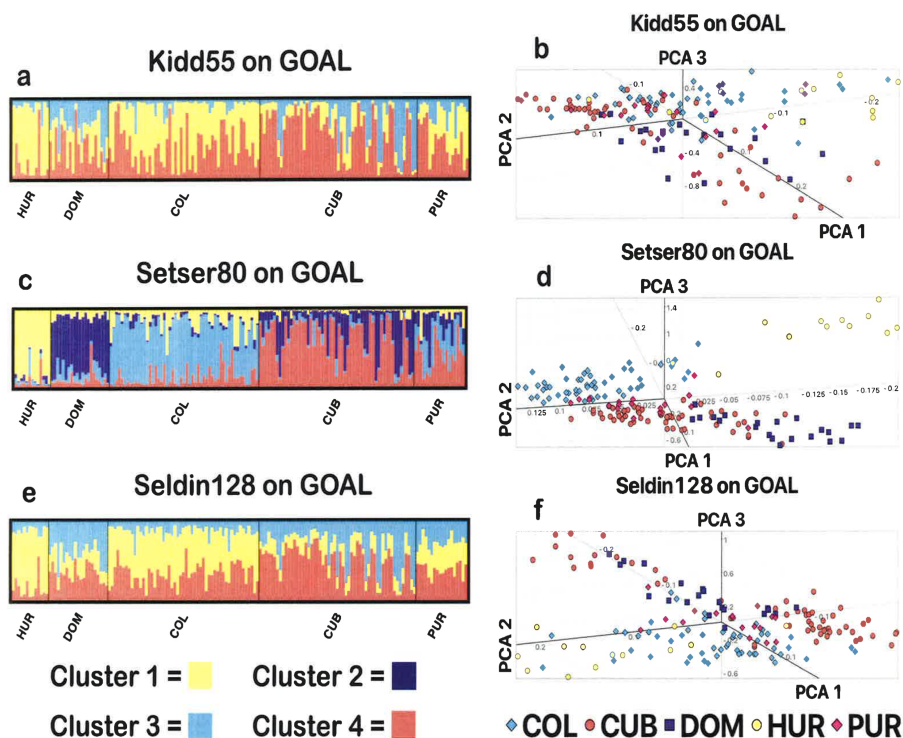
However, the gene chip used<sup>26</sup> was based on an early genome build (NCBI35/hg17) which did not have all the tag SNPs necessary (in comparison to the 1000G Project) to reliably impute ~30 of the SNPs from Seldin128<sup>9</sup> and 11 from Kidd55<sup>11</sup> for each individual. We assessed the accuracy of the imputation using the concordance tables provided by IMPUTE2; of the ~160 imputed SNPs from 20 chromosomes the mean concordance = 92.6% and range = 85.3% to 96.4%. Of the ~30 SNPs with missingness > 10%, there was no obvious pattern between missingness proportion and concordance. Despite multiple attempts with different

intervals, rs10954737 from the Seldin128<sup>9</sup> was unable to be imputed due to the lack of Panel 2 SNPs. Because STRUCTURE and PCA ignore missing data<sup>36, 37</sup>, the full Seldin128<sup>9</sup> and Kidd55<sup>11</sup> were used in these analyses. However, since the resampling approach to simulations is dependent upon the reliability of allele frequencies in our real data<sup>38</sup>, we applied the same <10% missingness filter used in the development of the Setser80; this resulted in 96 SNPs in the Seldin panel and 44 SNPs in the Kidd panel after imputation.

### **3.2.5 STRUCTURE**

We evaluated ancestry by the Bayesian model-based clustering method used in STRUCTURE v.2.3.4<sup>39</sup> to compare the self-reported to computer-determined (K) populations. We performed STRUCTURE analysis at K=2 to K=7 for each dataset/panel at 10 iterations each using the admixture model, no LOCPRIOR, 10,000 burn-in, and 100,000 Markov Chain Monte Carlo (MCMC) repetitions. The final STRUCTURE diagrams for each SNP panel were optimized and averaged through STRUCTURE Harvester<sup>40</sup>, CLUMPP<sup>41</sup>, and Distruct<sup>42</sup> to create the diagrams in Figure 1.

**Figure 1: Comparison to other panels**



**Figure 1: Comparison to Other Panels**

Each plot represents 160 unrelated GOAL individuals and their respective populations. Figures a, c, and e are STRUCTURE plots where each vertical line represents one person. Figures b, d, and f are PCA plots created through EIGENSOFT where the first three principal components are plotted. Figures a and b use the Kidd55 SNP panel (K=3), c and d use the Setser80 (K=4), and e and f use the Seldin128 (k=3). Abbreviations used: HUR = Honduras, DOM = Dominican Republic, COL = Colombia, CUB = Cuba, PUR = Puerto Rico, PCA = principal components analysis.

### 3.2.6 Principal Components Analysis (PCA)

We analyzed the Setser80, Seldin128<sup>9</sup>, and Kidd55<sup>11</sup> on the GOAL dataset by PCA using EIGENSOFT v.6.1.4<sup>43</sup> and plotted the first three eigenvectors. Genesis<sup>44</sup> was used for improved visualization of clustering as seen in Figure 1.

### **3.2.7 Linkage Disequilibrium (LD) Analysis**

Using the web-based tool LDmatrix<sup>45</sup>, we compared the Setser80 to the Seldin128<sup>9</sup> and Kidd55<sup>11</sup>, and the 46 Consensus SNPs described in a review article by Soundararajan et al.<sup>24</sup>. We used  $r^2 > 0.7$  as the threshold to evaluate whether any SNP in the Setser80 was in strong LD with SNP(s) from Seldin128<sup>9</sup> and Kidd55<sup>11</sup> (tested together) or the 46 Consensus SNPs appearing in more than 3 of 21 panels of AIMS<sup>24</sup>.

### **3.2.8 Modeling for the Prediction of Unknowns**

To model the data for BGA prediction of unknown individuals, we used a resampling approach based on calculated allele frequencies of the three SNP panels on each dataset<sup>38</sup>. We simulated a randomly mating population of 100-125 individuals within each country. Next, we assigned a genotype to individuals by generating a random number between 1 and 0 and comparing this number to the maf for the country at the specified locus. Any random number above the maf was assigned the major allele. All genotypes were created from 2 separate allele generations for each locus. The simulation of each population was performed at least 5 times for the GOAL and 1000G countries. The 7 Populations Combined dataset was created by merging the countries from the 1000G and GOAL simulations without regard to simulation number. We verified our model using a chi-square test for each panel and found the allele frequencies from the simulation sets were not significantly different from the true allele frequencies at  $\alpha = 0.05$  after Bonferroni correction.

### **3.2.9 Classification of Unknowns**

Snippet 2.5 app suite<sup>46</sup> is a web-based Naïve Bayes classifier, found here (<http://mathgene.usc.es/snippet/>), which calculates  $-\log(\text{likelihood})$  with leave-one-out cross-validation and multinomial logistic regression (MLR) options. Cross-validation divides a set of data into a

training set and a testing set, and rotates the samples until all samples have been in the testing set. Using the “Thorough analysis of population data with a custom Excel file” option, Snipper calculated likelihood ratios (LR) of *population vs. not the population* and selected the country that corresponded to the highest LR. MLR is similar to STRUCTURE<sup>39</sup>, which calculated genetic proportions of individuals (as percent admixture) instead of whole populations, and categorized individuals based on those probabilities. We used 100-125 micro-simulations (individuals) each from population for references and selected 10% of profiles from a separate set of micro-simulations to predict unknowns. We evaluated potential overlap of MLR classification using the confusion matrix and assessed the validity of our classification by sensitivity, specificity, and positive predictive value from the naïve Bayes classification of the actual 1000G genotypes (n=347; CLM=94, PUR=104, PEL=85, and MXL=64).

### **3.2.10 Ethical Approval and Informed Consent**

This research study using the Genomic Origins and Admixture in Latinos (GOAL) from Moreno-Estrada, A. *et al.* (2013)<sup>25</sup>, and the 1000 Genomes Project<sup>27</sup> datasets was approved under University of North Texas Health Science IRB 2013-201. As this manuscript only used pre-existing genetic data from Moreno-Estrada, A. *et al.* (2013)<sup>25</sup>, where their “Informed consent was obtained from all participants under approval by the University of Miami Institutional Review Board (study no. 20081175)”. The 1000 Genomes Project data was only included in the International Genome Sample Resource if the submission was in accordance with the Consent, Ethics Review and Sampling Process of the 1000 Genomes Project<sup>27</sup>.

### 3.3 Results

#### 3.3.1 Setser80 SNP Panel Evaluation

We evaluated the ability of a newly developed Hispanic AIMs panel (the Setser80) versus the Seldin128<sup>9</sup> and Kidd55<sup>11</sup> to separate heterogeneous Hispanic populations in the GOAL dataset (from Moreno-Estrada et al.<sup>25</sup>) using STRUCTURE<sup>39</sup> and principal components analysis (PCA). With the STRUCTURE<sup>39</sup> results, we applied the Evanno method<sup>47</sup> which optimized the computer-determined (K) populations; the highest likelihood for the Setser80 was at K=4 while Seldin128<sup>9</sup> and Kidd55<sup>11</sup> were optimized at K=3 (Figure 1a, 1c, 1e). The genetic proportions from STRUCTURE<sup>39</sup> indicated that the Setser80 clearly separates HUR (Cluster 1 = 0.8290), DOM (Cluster 2 = 0.6976), and COL (Cluster 3 = 0.6562) (Table 1); but CUB (Cluster 2 = 0.2892, Cluster 4 = 0.6125) and PUR (Cluster 2 = 0.2048, Cluster 4 = 0.4145) remain indistinguishable (Figure 1c). Using the genetic proportions from STRUCTURE<sup>39</sup> for the Seldin128<sup>9</sup> and Kidd55<sup>11</sup> panels, HUR and COL separated predominately into Cluster 1 (HUR: Seldin128<sup>9</sup> = 0.7274, Kidd55<sup>11</sup> = 0.7258)(COL: Seldin128<sup>9</sup> = 0.5370, Kidd55<sup>11</sup> = 0.5311) (Table 1), but the remaining populations did not separate into distinct clusters.

We performed a principal components analysis (PCA) for the AIMs panels in the GOAL population (Figure 1b, 1d, 1f). In the PCA of the Setser80, HUR clearly separated across PC1 and PC2, DOM separated from HUR across PC2, and COL separated from HUR across PC1 and from DOM across PC2, which occupies three separate quadrants of the PCA (Figure 1d). Seldin128<sup>9</sup> PCA showed HUR and COL separated together but apart from the other populations across PC1, and CUB and DOM separated together along PC2 (Figure 1f). The Kidd55<sup>11</sup> performed poorly in PCA (Figure 1b), not forming recognizable clusters, consistent with the genetic proportions generated in STRUCTURE<sup>39</sup> (Figure 1a)(Table 1). The Setser80 was able to

differentiate HUR, DOM, and COL by the two different algorithms underlying STRUCTURE<sup>39</sup> and PCA.

### **3.3.2 Classification of Unknowns**

Based on the GOAL<sup>25</sup> and 1000 Genomes Project<sup>27</sup> (1000G) allele frequencies, we modeled populations to determine classification accuracy using the Snipper 2.5<sup>46</sup> app suite. Snipper uses naïve Bayesian likelihood ratios and multinomial logistic regression (MLR) prediction of unknowns via  $-\log(\text{likelihood})$ <sup>46</sup>. Despite the different algorithms, both analyses had similar results.

As expected, the Setser80 had the highest overall accuracy across the three panels in the simulated GOAL dataset (98.4%) by naïve Bayesian classification implemented via leave-one-out cross-validation. Additionally, the Setser80 achieved 90% accuracy in the 1000G dataset and 91.5% in the 7 Populations Combined dataset, both of which include populations not involved in our SNP ascertainment (Table 2). In the latter, the Setser80 panel (98%) and the Seldin panel (98.8%) achieved approximately equal accuracy in PEL, a population on which the Setser80 was not trained. In the 1000G simulations, the Seldin panel was more accurate overall (92.4%) in comparison to the Setser80 (90%).

Naïve Bayes analysis of the actual 1000G genotypes revealed the Setser80 had the highest specificity in CLM (98.4%), the highest sensitivity in MXL (84.4%), and similar specificity in PUR (85.2%) and PEL (97.7%) in comparison to the Seldin (86.8%, 95.4%) and Kidd (85.2%, 94.7%) panels (Table 3). In all three SNP panels, the micro-simulations underestimated the positive predictive value of CLM. The positive predictive value of Setser80 for PUR (simulated = 69.8%, real = 70.2%) and PEL (simulated = 91.8%, real = 89.8%) was concordant between the simulated and real data where it was either under or overestimated by

the Seldin and Kidd panels. Both the Setser80 (simulated = 59.3%, real = 36.7%) and the Seldin (simulated = 80.1%, real 54.1%) panels overestimated positive predictive value in MXL while the Kidd panel values were concordant between the simulations (47.3%) and real genotypes (45.7%).

Utilizing the MLR algorithm, Setser80 had the highest accuracy in GOAL and 7 Populations Combined (99% and 93.2%, respectively); the Setser80 and Seldin panel had equal accuracy in 1000G (93.8%); and the Kidd panel had 80.5% in GOAL, 71.4% in 7 Populations Combined, and 82.2% overall in 1000G (Table 3). As expected, HUR achieved >95% accuracy in the Setser80 and the Seldin panel across all datasets. Surprisingly, PEL also achieved >95% and MXL >90% accuracies using the Setser80, although the Setser80 had not been trained on these populations.

Despite performing best overall, the Setser80 did misclassify COL 22.5% of the time in the 7 Populations Combined dataset (Supplemental Table S3.3. When it misclassified COL, the individual was classified as MXL 77.8% and PUR 22.2% of the time. Conversely, even though MXL classified correctly 90% of the time, when individuals were misclassified they were misclassified as COL 100% of the time. In comparison, the Seldin panel misclassified COL 17.5% of the time spread across four countries, primarily into PUR (10%). The Kidd panel exhibited a similar trend where COL misclassified into five countries: PUR (15%), MXL (10%), HUR (7.5%), CUB (7.5%), and DOM (2.5%) in addition to one individual which could not be classified. When MXL was misclassified using the Kidd panel, it misclassified into PEL (7.5%), HUR (5%), and COL (5%). Additionally, the Kidd panel had high misclassification of HUR into MXL (20%), COL (15%), and PUR (7.5%).

### 3.4 Discussion

We report a panel of 80 AIMs for Hispanic BGA classification using Weir & Cockerham's estimator<sup>32</sup> of Wright's  $F_{ST}$ <sup>31</sup>. Honduras (HUR) and DOM emerged first in STRUCTURE<sup>39</sup> and PCA, followed by COL at K=4, which separated from CUB & PUR, indicating three distinct populations (Table 1). Based on the allele frequencies, we created a series of micro-simulations to compare the BGA classification of the Setser, Seldin, and Kidd panels. Overall, the Setser80 outperformed the Seldin and Kidd panels in naïve Bayesian classification and MLR classification accuracies in the GOAL dataset (naïve Bayes = 98.4%, MLR = 99%) and the 7 Populations Combined (naïve Bayes = 91.5%, MLR = 93.2%). Notably, PEL and MXL were classified with >95% and >80% accuracy, respectively, indicating the Setser80 panel is portable into other Hispanic datasets and populations.

Many panels have sought country-level ancestry determination, using a variety of SNP ascertainment methods<sup>19, 21, 27, 46</sup>. Continentally, the EUROFORGEN Global AIMs<sup>12</sup> and the Kidd55<sup>11</sup> panel used allele frequency differentials ( $\delta$ ). Within a country, the United States HapMap 3 populations<sup>20</sup> used PCA with receiver operating characteristics curve (ROC)<sup>19</sup>, and the Mexican mestizos panel used nested subsets with high SNP weights followed by the lowest number of SNPs with the highest PC1<sup>21</sup>. Similar to Kidd et al.<sup>11</sup>, we prioritized SNPs that distinguished populations with lower mean  $F_{ST}$  per country. However, we focused on differentiating Hispanic instead of continental populations. Kosoy et al.<sup>9</sup> (Seldin128) also concentrates on continental differentiation, but they also evaluated their AIMs on African American, Puerto Rican, and Mexican/Mexican American populations.

We used the Snipper 2.5 app suite<sup>46</sup> that provided two classification methods: a naïve Bayesian classifier and MLR<sup>48</sup>. This web-based classifier was designed for classification of

externally visible characteristics<sup>49-53</sup> and ancestry<sup>12, 54-57</sup>, particularly in forensics. Snipper has successfully analyzed admixed South American populations<sup>50, 58, 59</sup>, similar to those used here.

The classification accuracy of the Seldin and Kidd panels is due to both the composition of their SNP ascertainment datasets and the size of the panels. The Seldin panel (96.2%, 96.3%) was more accurate in MXL than the Setser80 (83.4%, 89.8%) in the 7 Populations Combined and 1000G datasets, respectively. Its success is likely because 199 of their 825 samples were from admixed Latin American and Amerindian individuals (Mexico and Puerto Rico especially)<sup>9</sup>. The Kidd<sup>11</sup> panel emphasized capturing diversity by using 63 global populations<sup>11</sup> including seven isolated Amerindian populations; they continue to add more populations via ALFRED<sup>60</sup>. The size of the Kidd panel and the ratio of SNPs to the number of samples (Kidd55 = 55 SNPs / 3071 samples = 0.0179; Seldin128 = 128 SNPs / 825 samples = 0.1552) suggest the number of SNPs, rather than SNP ascertainment population size, is the higher contributing factor to population differentiation. However, the number of individuals per population may also be a factor.

Our study's limitations include: genechip design, sample size and its effect on allele frequencies, the use of a static model, and missingness. The GOAL<sup>25</sup> study genechip<sup>26</sup> was built on 270 African (YRI), Caucasian (CEU), and East Asian population (CHB and JPT) samples from HapMap 1<sup>61</sup>, without any Amerindian component. Although, our SNP ascertainment dataset was small it was not inconsistent with other studies<sup>11, 18, 20</sup> where the larger overall size was coupled with small sub-populations. Therefore, we combined the GOAL<sup>25</sup> dataset with the 1000 Genomes Admixed American dataset (n=347)<sup>27</sup>, merging COL with CLM (n=147) and PUR with PUR (n=122) due to negligible allele frequency differences, to create the 7 Populations Combined.

The design of the Setser80 is based on the balance of the countries via country attributable mean  $F_{ST}$  and selection of SNPs with  $LD < 0.7$ . Using a dilution series of 234 to 44 SNPs, we evaluated the effect of panel size on classification accuracy in relation to Seldin and Kidd sized panels and found 80 SNPs to be sufficient. Therefore we chose 80 SNPs from 247 candidates by selecting SNPs such that ~20% could be attributed to each country. It is possible that other panels informative of Hispanic ancestry could be selected from the same candidates, but testing multiple different panels was beyond the scope of this study. Residual LD is possible despite our threshold where four pairs of SNPs had  $r^2 > 0.5$ ; however, removing one of each pair and classifying two separate 76 SNP subsets had negligible effect on classification accuracy via naïve Bayes (Supplemental Table S3.4) or MLR (Supplemental Table S3.5). By treating these loci as independent, we may underestimate accuracy as Kidd et al. 2013 has shown that diplotypes are effective predictors of ancestry<sup>62</sup>.

We used micro-simulations in this study in order to normalize the size of each population and expand the analysis to seven Hispanic populations instead of the four publicly available through the 1000 Genomes Project<sup>27</sup>. Although real genotypes would have been preferable, widely variable population sizes could disproportionately affect the classification accuracy for smaller populations, as may have been the case with the real MXL genotypes. Our analysis of additional populations is a more realistic representation of the challenges of a more granular classification of heterogeneous populations. Forensic labs may not have access to a sizeable Hispanic database of individuals from multiple different countries; therefore, we simulated datasets based on readily available allele frequencies from multiple sources. By doing so, we have allowed MXL to misclassify into HUR which otherwise do not exist within the same dataset.

Additionally, our use of a static model for BGA determination may have overestimated classification success; despite reasonable success by other research groups<sup>63</sup>. Finally, our imputation of the Seldin128<sup>9</sup> and Kidd55<sup>11</sup> into the GOAL<sup>25</sup> dataset required removal of ~30 loci to comply with the Setser80 QC filters. Missingness was not detrimental here because STRUCTURE disregards it<sup>36, 37</sup>, and at 10% MLR is robust<sup>36</sup>. Alternatively, some missingness in micro-simulations may approximate the degraded forensic samples<sup>64</sup>.

Our findings indicate that the Setser80 can predict BGA of individuals of presumed Hispanic origin with high confidence. By selecting additional SNPs attributed to countries with lower average country attributable  $F_{ST}$  (COL, CUB, and PUR) to create the panel, the Setser80 had similar accuracy overall in GOAL<sup>25</sup> and 7 Populations Combined. The Setser80 is robust as it clusters well with Bayesian model-based clustering and PCA, and classifies equally well in naïve Bayes classification and MLR. The Setser80 is portable and, to our knowledge, is the first BGA AIMs panel specifically for the Caribbean and surrounding mainland countries. In comparison to Seldin128<sup>9</sup>, Kidd55<sup>11</sup>, and 46 Consensus SNPs<sup>24</sup>, our 80 AIMs for Hispanic BGA is unique, both exact and by linkage disequilibrium. Therefore, it is our intention that the Setser80 be integrated into a future Western Hemisphere panel.

### 3.5 References

- 1) Gao, C. *et al.* A comprehensive analysis of common and rare variants to identify adiposity loci in Hispanic Americans: The IRAS family study (IRASFS). *PLoS ONE*. **10** (11) e0134649; 10.1371/journal.pone.0134649 (2015 November 1).
- 2) Phillips, C. Forensic genetic analysis of bio-geographical ancestry. *Forensic Sci Int-Gen.* **18**, 49-65 (2015).

- 3) Burkart, K. M. *et al.* A genome-wide association study in Hispanics/Latinos identifies novel signals for lung function – The Hispanic Community Health Study/Study of Latinos. *Am J Resp Crit Care Med.* **198** (2), 208–219 (2018 July 15).
- 4) Manichaikul, A., *et al.* Population structure of Hispanics in the United States: The multi-ethnic study of atherosclerosis. *PLoS Genet.* **8** (4), e1002640; 10.1371/journal.pgen.1002640 (2012 April).
- 5) MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Res.* **45** (D1), D896-D901 (2017 January 1).
- 6) Norris, E. T. *et al.* Genetic ancestry, admixture and health determinants in Latin America. *BMC Genomics.* **19** (Suppl 8), Article 861 (2018 December).
- 7) Ambers, A. D. *et al.* Comprehensive forensic genetic marker analysis for accurate human remains identification using massively parallel DNA sequencing. *BMC Genomics.* **17** (Suppl 9), Article 750 (2016 October 17).
- 8) Lorente, J. A. Trafficking in human beings: modern slavery. EndSlavery. Workshop 2-3, November 2013. URL [http://www.endslavery.va/content/endslavery/en/publications/scripta\\_varia\\_122/lorente.html](http://www.endslavery.va/content/endslavery/en/publications/scripta_varia_122/lorente.html) (2019).
- 9) Kosoy, R. *et al.* Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat.* **30** (1), 69-78 (2009 January).
- 10) Galanter, J. M. *et al.* Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genet.* **8** (3), e1002554; 10.1371/journal.pgen.1002554 (2012 March).
- 11) Kidd, K. K. *et al.* Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci Int-Gen.* **10** (1), 23-32 (2014 May).
- 12) Phillips, C. *et al.* Building a forensic ancestry panel from the ground up: the EUROFORGEN Global AIM-SNP set. *Forensic Sci Int-Gen.* **11** (1), 13-25 (2014 July).
- 13) Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science.* **343** (6172), 747-751 (2014 February 14).
- 14) Elhaik, E. *et al.* Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat Commun.* **5**, 3513; 10.1038/ncomms4513 (2014 April 29).

- 15) Marcheco-Teruel, B. *et al.* Cuba: Exploring the history of admixture and the genetic basis of pigmentation using autosomal and uniparental markers. *PLoS Genet.* **10** (7), e1004488; 10.1371/journal.pgen.1004488 (2014 July 14).
- 16) Fortes-Lima, C. *et al.* Exploring Cuba's population structure and demographic history using genome-wide data. *Sci Rep.* **8** (1), 11422; 10.1038/s41598-018-29851-3 (2018 December 1).
- 17) Jia, J., Wei, Y., Qin, C., Hu, L., Wan, L. & Li, C. Developing a novel panel of genome-wide ancestry informative markers for bio-geographical ancestry estimates. *Forensic Sci Int-Gen.* **8**, 187–194 (2014).
- 18) Tian, C. *et al.* Analysis of East Asia genetic substructure using genome-wide SNP arrays. *PLoS ONE.* **3** (12), e3862; 10.1371/journal.pone.0003862 (2008 December 5).
- 19) Zeng, X., Chakraborty, R., King, J. L., LaRue, B., Moura-Neto, R. S. & Budowle, B. Selection of highly informative SNP markers for population affiliation of major US populations. *Int J Legal Med.* **130** (2), 341–352 (2016 March).
- 20) Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature.* **467** (7311), 52-58 (2010 September 2).
- 21) Huerta-Chagoya, A. *et al.* A panel of 32 AIMs suitable for population stratification correction and global ancestry estimation in Mexican mestizos. *BMC Genetics.* **20** (1), 5; 10.1186/s12863-018-0707-7 (2019 January 8).
- 22) Elhaik, E. *et al.* The GenoChip: A new tool for genetic anthropology. *Genome Biol Evol.* **5** (5), 1021-1031 (2013).
- 23) Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature.* **519** (7543), 309-314 (2015 March 19).
- 24) Soundararajan, U., Yun, L., Shi, M. & Kidd, K. K. Minimal SNP overlap among multiple panels of ancestry informative markers argues for more international collaboration. *Forensic Sci Int-Gen.* **23**, 25-32 (2016 July 1).
- 25) Moreno-Estrada, A. *et al.* Reconstructing the population genetic history of the Caribbean. *PLoS Genet.* **9** (11), e1003925; 10.1371/journal.pgen.1003925 (2013 November 14).
- 26) McCarroll, S. A. *et al.* Integrated detection and population genetic analysis of SNPs and copy number variation. *Nat Genet.* **40** (10), 1166-1174 (2008 October).
- 27) Auton, A. *et al.* A global reference for human genetic variation. *Nature.* **526** (7571), 68–74 (2015 October 1).

- 28) Kent, W. J. *et al.* UCSC Genome Browser: The human genome browser at UCSC. *Genome Res.* **12** (6), 996-1006 (2002 June). URL <http://genome.ucsc.edu>.
- 29) Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* **81** (3), 559-575 (2007 September).
- 30) Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M. & Lee, J. J. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience.* **4** (1), Article 7, (2015 February 25).
- 31) Wright, S. The genetical structure of populations. *Ann Eugen.* **15** (4), 323-354 (1951 March).
- 32) Weir, B. S. & Cockerham, C. C. Estimation of gene flow from F-statistics. *Evolution.* **47** (3), 855-863 (1993).
- 33) Purcell, S. & Chang, C. PLINK 1.9. URL <http://www.cog-genomics.org/plink/1.9/>.
- 34) Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5** (6), e1000529; 10.1371/journal.pgen.1000529 (2009 June).
- 35) Howie, B. & Marchini, J. Instructions for IMPUTE version 2. (2009 June 18). URL [https://mathgen.stats.ox.ac.uk/impute/impute\\_v2\\_instructions.pdf](https://mathgen.stats.ox.ac.uk/impute/impute_v2_instructions.pdf)
- 36) Cheung, E. Y. Y., Gahan, M. E. & McNeven, D. Prediction of biogeographical ancestry from genotype: A comparison of classifiers. *Int J Legal Med.* **131** (4), 901-912 (2017 July 1).
- 37) Pritchard, J. K., Wen, X. & Falush, D. Documentation for STRUCTURE software: Version 2.3. (2010 February 2).
- 38) Yuan, X., Miller, D. J., Zhang, J., Herrington, D. & Wang, Y. An overview of population genetic data simulation. *J Comput Biol.* **19** (1), 42-54 (2012 January 1).
- 39) Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics.* **155** (2), 945-959 (2000 June).
- 40) Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour.* **4** (2), 359-361 (2012).
- 41) Jakobsson, M. & Rosenberg, N. A. CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics.* **23** (14), 1801-1806 (2007 July 15).

- 42) Rosenberg, N. A. Distruct: A program for the graphical display of population structure. *Mol Ecol Notes*. **4** (1), 137-138 (2004 March).
- 43) Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet*. **2** (12), 2074-2093; 10.1371/journal.pgen.0020190 (2006 December 22).
- 44) Buchmann, R.W. Genesis: Copyright © 2014, University of the Witwatersrand.
- 45) National Cancer Institute, Division of Cancer Epidemiology & Genetics. LD Matrix. URL <https://ldlink.nci.nih.gov/?tab=ldmatrix> (2019).
- 46) Phillips, C. *et al.* Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int-Gen*. **1** (3-4), 273–280 (2007 December).
- 47) Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol Ecol*. **14** (8), 2611-2620 (2005 July).
- 48) McNevin, D. *et al.* An assessment of Bayesian and multinomial logistic regression classification systems to analyse admixed individuals. *Forensic Sci Int-Gen*. Supplement Series **4** (1), e63-e64; 10.1016/j.fsigss.2013.10.032 (2013).
- 49) Ruiz, Y. *et al.* Further development of forensic eye color predictive tests. *Forensic Sci Int-Gen*. **7** (1), 28-40 (2013 January).
- 50) Freire-Aradas, A. *et al.* Exploring iris colour prediction and ancestry inference in admixed populations of South America. *Forensic Sci Int-Gen*. **13**, 3-9 (2014 November).
- 51) Maroñas, O. *et al.* Development of a forensic skin colour predictive test. *Forensic Sci Int-Gen*. **13**, 34-44 (2014 November).
- 52) Söchtig, J. *et al.* Exploration of SNP variants affecting hair colour prediction in Europeans. *Int J Legal Med*. **129** (5), 963-975 (2015 September).
- 53) Pośpiech, E. *et al.* The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction. *Forensic Sci Int-Gen*. **11**, 64-72 (2014 July).
- 54) Fondevila, M. *et al.* Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies. *Forensic Sci Int-Gen*. **7** (1), 63-74 (2013 January).
- 55) Pereira, R. *et al.* Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. *PLoS ONE*. **7** (1), e29684; 10.1371/journal.pone.0029684 (2012 January 17).

- 56) De la Puente, M. *et al.* The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs. *Forensic Sci Int-Gen.* **22**, 81-88 (2016 May).
- 57) Eduardoff, M. *et al.* Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM™. *Forensic Sci Int-Gen.* **23**, 178-189 (2016 July 1).
- 58) Heinz, T. *et al.* Ancestry analysis reveals a predominant Native American component with moderate European admixture in Bolivians. *Forensic Sci Int-Gen.* **7** (5), 537-542 (2013).
- 59) Taboada-Echalar, P. *et al.* The genetic legacy of the pre-colonial period in contemporary Bolivians. *PLoS ONE.* **8** (3), e58980; 10.1371/journal.pone.0058980 (2013 March).
- 60) Rajeevan, H., Soundararajan, U., Kidd, J., R., Pakstis, A. & Kidd, K. K. ALFRED: An allele frequency resource for research and teaching. *Nucleic Acids Res.* **40** (D1), D1010-D1015 (2012 January).
- 61) The International HapMap Consortium. The International HapMap Project. *Nature.* **426** (6968), 789-796 (2003 December 18).
- 62) Kidd, K. K. *et al.* Microhaplotype loci are a powerful new type of forensic marker. *Forensic Sci Int-Gen.* Supp Series 4 (1), e123–e124; 10.1016/j.fsigs.2013.10.063 (2013).
- 63) Kusev, P., van Schaik, P., Tsaneva-Atanasova, K., Juliusson, A. & Chater, N. Adaptive anchoring model: How static and dynamic presentations of time series influence judgment predictions. *Cogn Sci.* **42** (1), 77-102 (2018 January).
- 64) Butler, J. M. (2<sup>nd</sup> ed.) Forensic DNA typing: Biology, technology, and genetics of STR markers. (Elsevier, 2005).

### 3.6 Acknowledgements

Funding support for the Genomic Origins and Admixture in Latinos (GOAL) Study was provided by the National Institute of General Medical Sciences (1R01GM090087). Additional support for sample collection was provided by a grant from the Stanley J. Glaser Foundation and the Dr. John T. Macdonald Foundation Department of Human Genetics. The dataset used for the

analyses described in this manuscript was obtained from dbGaP through accession number phs000750.v1.p1.

The authors would like to thank the late Dr. Arthur Eisenberg for the inspiration behind this project based on the needs of the Center for Human Identification and DNA ProKids. We would also like to thank the late Dr. Ranajit Chakraborty who was instrumental in the design of this research. We thank Dr. Carlos Bustamante for making the GOAL dataset available via dbGaP. Dr. Xiangpei Zeng helped with STRUCTURE and Dr. Frank Wendt helped with access and use of data from 1000 Genomes. Dr. Gita Pathak helped in many small but significant ways in discussing concepts and troubleshooting software.

### **3.7 Author Contributions**

C.H.S. designed the project alongside R.C., performed the analyses, interpreted results alongside J.V.P. and R.B., troubleshoot with input from N.P., and prepared the manuscript. D.S.C. has provided substantial guidance in experimental design, interpretation, troubleshooting, and especially crafting research into a publishable manuscript.

### **3.8 Additional Information**

**Competing Interests:** The authors declare no competing interests.

### 3.9 Tables

**Table 1: Genetic Proportions from STRUCTURE**

| Panel              | Population | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Individuals |
|--------------------|------------|-----------|-----------|-----------|-----------|-------------|
| Setser80<br>(K=4)  | HUR        | 0.8290    | 0.0387    | 0.0647    | 0.0676    | 13          |
| Setser80<br>(K=4)  | DOM        | 0.0811    | 0.6976    | 0.1147    | 0.1067    | 21          |
| Setser80<br>(K=4)  | COL        | 0.1601    | 0.0474    | 0.6562    | 0.1365    | 53          |
| Setser80<br>(K=4)  | CUB        | 0.0348    | 0.2892    | 0.0634    | 0.6125    | 55          |
| Setser80<br>(K=4)  | PUR        | 0.0836    | 0.2048    | 0.2969    | 0.4145    | 18          |
| Seldin128<br>(K=3) | HUR        | 0.7274    | 0.1155    | 0.1570    | N/A       | 13          |
| Seldin128<br>(K=3) | DOM        | 0.2296    | 0.4283    | 0.3422    | N/A       | 21          |
| Seldin128<br>(K=3) | COL        | 0.5370    | 0.1280    | 0.3349    | N/A       | 53          |
| Seldin128<br>(K=3) | CUB        | 0.1672    | 0.3507    | 0.4822    | N/A       | 55          |
| Seldin128<br>(K=3) | PUR        | 0.3415    | 0.2728    | 0.3860    | N/A       | 18          |
| Kidd55<br>(K=3)    | HUR        | 0.7258    | 0.1077    | 0.1664    | N/A       | 13          |
| Kidd55<br>(K=3)    | DOM        | 0.2664    | 0.3548    | 0.3788    | N/A       | 21          |
| Kidd55<br>(K=3)    | COL        | 0.5311    | 0.0690    | 0.4001    | N/A       | 53          |
| Kidd55<br>(K=3)    | CUB        | 0.1723    | 0.2528    | 0.5749    | N/A       | 55          |
| Kidd55<br>(K=3)    | PUR        | 0.3907    | 0.1705    | 0.4389    | N/A       | 18          |

**Table 1: Genetic Proportions from STRUCTURE**

Each vertical line in a STRUCTURE diagram represents one individual, and the values listed here correspond to the genetic proportions of each of “K” computer determined populations, represented as colors in the diagram. The Setser80 categorized genetic proportions of samples into four computer-determined populations (K=4). The Seldin128 and Kidd55 categorized genetic proportions into three computer-determined populations (K=3).

**Table 2: Naïve Bayesian Classification Accuracy**

| <b>Panel<br/>(Dataset)</b>   | <b>HUR</b>       | <b>DOM</b>       | <b>COL</b>       | <b>CUB</b>       | <b>PUR</b>       | <b>PEL</b>        | <b>MXL</b>        | <b>Overall</b> |
|------------------------------|------------------|------------------|------------------|------------------|------------------|-------------------|-------------------|----------------|
| <b>Setser80<br/>(GOAL)</b>   | 100%<br>(±0%)    | 96.8%<br>(±2.5%) | 99.4%<br>(±0.5%) | 96.8%<br>(±2.8%) | 99%<br>(±0.7%)   | N/A               | N/A               | 98.4%          |
| <b>Seldin96<br/>(GOAL)</b>   | 99.2%<br>(±0.4%) | 89.6%<br>(±3%)   | 78.4%<br>(±4.2%) | 76%<br>(±3.3%)   | 90.8%<br>(±1.8%) | N/A               | N/A               | 87.9%          |
| <b>Kidd44<br/>(GOAL)</b>     | 88.4%<br>(±3.4%) | 78.6%<br>(±4.1%) | 67.6%<br>(±4%)   | 66.2%<br>(±5.3%) | 68%<br>(±7.3%)   | N/A               | N/A               | 73.8%          |
| <b>Setser80<br/>(1000G)</b>  | N/A              | N/A              | 81.9%<br>(±2.7%) | N/A              | 90.4%<br>(±2.1%) | 98.1%<br>(±0.9%)  | 89.8%<br>(±3%)    | 90%            |
| <b>Seldin96<br/>(1000G)</b>  | N/A              | N/A              | 84.2%<br>(±3.6%) | N/A              | 89.8%<br>(±4.9%) | 99.4%<br>(±0.7%)  | 96.3%<br>(±1.5%)  | 92.4%          |
| <b>Kidd44<br/>(1000G)</b>    | N/A              | N/A              | 63.2%<br>(±1.9%) | N/A              | 75.84%<br>(±3%)  | 91.84%<br>(±2.7%) | 85.28%<br>(±3.3%) | 79.00%         |
| <b>Setser80<br/>(7 Pops)</b> | 98.4%<br>(±0.9%) | 97.4%<br>(±1.7%) | 77.6%<br>(±8.2%) | 95.8%<br>(±1.9%) | 89.8%<br>(±2.9%) | 98%<br>(±1%)      | 83.4%<br>(±3.3%)  | 91.5%          |
| <b>Seldin96<br/>(7 Pops)</b> | 85%<br>(±2.5%)   | 84.4%<br>(±3.1%) | 79.8%<br>(±4.6%) | 68.8%<br>(±3.1%) | 79.6%<br>(±7%)   | 98.8%<br>(±0.8%)  | 96.2%<br>(±0.8%)  | 84.7%          |
| <b>Kidd44<br/>(7 Pops)</b>   | 67.8%<br>(±7.8%) | 83.2%<br>(±5.1%) | 59%<br>(±4.4%)   | 61.2%<br>(±4.3%) | 56.4%<br>(±2.1%) | 91.4%<br>(±1.1%)  | 78.6%<br>(±4.6%)  | 71.1%          |

**Table 2: Naïve Bayesian Classification Accuracy**

Comparison of the nine possible combinations of each of three simulated datasets on each of three SNP panels and their naïve Bayesian classification accuracy for each population. Reported as percent accuracy with two-tailed standard deviations listed in parentheses (). Abbreviations used: GOAL = Genomic Origins and Admixture in Latinos, 1000G = 1000 Genomes Project, 7 Pops = 7 Populations Combined, COL = Colombia, CUB = Cuba, DOM = Dominican Republic, HUR = Honduras, PUR = Puerto Rico, PEL = Peru from Lima, and MXL = Mexicans living in Los Angeles. Both Colombian populations from GOAL and 1000G are listed in this table as “COL”.

**Table 3: Positive predictive values from naïve Bayes analysis**

| Known Origin | SNP Panel | 5 sets of 500 micro-simulations |          |         | 347 real 1000G genotypes |          |         |
|--------------|-----------|---------------------------------|----------|---------|--------------------------|----------|---------|
|              |           | Sen. (%)                        | Spe. (%) | PPV (%) | Sen. (%)                 | Spe. (%) | PPV (%) |
| CLM          | Setser80  | 81.9%                           | 70.1%    | 47.8%   | 17.0%                    | 98.4%    | 80.0%   |
|              | Seldin96  | 84.2%                           | 77.9%    | 55.9%   | 55.3%                    | 90.9%    | 69.3%   |
|              | Kidd44    | 63.2%                           | 49.9%    | 29.6%   | 51.1%                    | 83.8%    | 53.9%   |
| PUR          | Setser80  | 90.4%                           | 86.9%    | 69.8%   | 81.7%                    | 85.2%    | 70.2%   |
|              | Seldin96  | 89.8%                           | 80.5%    | 60.6%   | 89.4%                    | 86.8%    | 74.4%   |
|              | Kidd44    | 75.8%                           | 51.7%    | 34.4%   | 71.2%                    | 85.2%    | 67.3%   |
| PEL          | Setser80  | 98.1%                           | 97.1%    | 91.8%   | 62.4%                    | 97.7%    | 89.8%   |
|              | Seldin96  | 99.4%                           | 98.9%    | 96.9%   | 87.1%                    | 95.4%    | 86.0%   |
|              | Kidd44    | 91.8%                           | 90.4%    | 76.1%   | 75.3%                    | 94.7%    | 82.1%   |
| MXL          | Setser80  | 89.8%                           | 79.5%    | 59.3%   | 84.4%                    | 67.1%    | 36.7%   |
|              | Seldin96  | 96.3%                           | 92.0%    | 80.1%   | 51.6%                    | 90.1%    | 54.1%   |
|              | Kidd44    | 85.3%                           | 68.3%    | 47.3%   | 50.0%                    | 86.6%    | 45.7%   |

**Table 3: Positive predictive values from naïve Bayes analysis.** Sensitivity, specificity, and positive predictive values from naïve Bayes leave-one-out cross-validation for the average of five sets of 500 micro-simulations (left) and n=347 actual 1000G genotypes (right). Micro-simulations were generated based on the allele frequencies from the 1000G dataset only. Abbreviations used: Sen. = sensitivity, Spe. = specificity, PPV = positive predictive value, CLM = Colombia from Medellin, PUR = Puerto Rico, PEL = Peru from Lima, and MXL = Mexicans living in Los Angeles.

**Table 4: MLR Classification Accuracy**

| <b>Panel<br/>(Dataset)</b>   | <b>HUR</b>               | <b>DOM</b>                | <b>COL</b>                | <b>CUB</b>                | <b>PUR</b>                | <b>PEL</b>               | <b>MXL</b>               | <b>Overall</b> |
|------------------------------|--------------------------|---------------------------|---------------------------|---------------------------|---------------------------|--------------------------|--------------------------|----------------|
| <b>Setser80<br/>(GOAL)</b>   | 100%<br>( $\pm 0\%$ )    | 100%<br>( $\pm 0\%$ )     | 100%<br>( $\pm 0\%$ )     | 97.5%<br>( $\pm 5\%$ )    | 97.5%<br>( $\pm 5\%$ )    | N/A                      | N/A                      | 99%            |
| <b>Seldin96<br/>(GOAL)</b>   | 97.5%<br>( $\pm 5\%$ )   | 95%<br>( $\pm 5.8\%$ )    | 85%<br>( $\pm 12.9\%$ )   | 90%<br>( $\pm 11.5\%$ )   | 95%<br>( $\pm 5.8\%$ )    | N/A                      | N/A                      | 92.5%          |
| <b>Kidd44<br/>(GOAL)</b>     | 92.5%<br>( $\pm 9.6\%$ ) | 90%<br>( $\pm 0\%$ )      | 75%<br>( $\pm 17.3\%$ )   | 72.5%<br>( $\pm 15\%$ )   | 72.5%<br>( $\pm 9.6\%$ )  | N/A                      | N/A                      | 80.5%          |
| <b>Setser80<br/>(1000G)</b>  | N/A                      | N/A                       | 90.4%<br>( $\pm 7.4\%$ )  | N/A                       | 90.4%<br>( $\pm 7.4\%$ )  | 100%<br>( $\pm 0\%$ )    | 94.2%<br>( $\pm 7.4\%$ ) | 93.8%          |
| <b>Seldin96<br/>(1000G)</b>  | N/A                      | N/A                       | 94.2%<br>( $\pm 3.8\%$ )  | N/A                       | 88.5%<br>( $\pm 7.7\%$ )  | 100%<br>( $\pm 0\%$ )    | 92.3%<br>( $\pm 6.3\%$ ) | 93.8%          |
| <b>Kidd44<br/>(1000G)</b>    | N/A                      | N/A                       | 76.9%<br>( $\pm 8.9\%$ )  | N/A                       | 76.9%<br>( $\pm 6.3\%$ )  | 92.3%<br>( $\pm 8.9\%$ ) | 82.7%<br>( $\pm 9.7\%$ ) | 82.2%          |
| <b>Setser80<br/>(7 Pops)</b> | 95%<br>( $\pm 5.8\%$ )   | 97.5%<br>( $\pm 5\%$ )    | 77.5%<br>( $\pm 9.6\%$ )  | 100%<br>( $\pm 0\%$ )     | 92.5%<br>( $\pm 9.6\%$ )  | 100%<br>( $\pm 0\%$ )    | 90%<br>( $\pm 8.2\%$ )   | 93.2%          |
| <b>Seldin96<br/>(7 Pops)</b> | 100%<br>( $\pm 0\%$ )    | 82.5%<br>( $\pm 20.6\%$ ) | 82.5%<br>( $\pm 12.6\%$ ) | 85%<br>( $\pm 17.3\%$ )   | 67.5%<br>( $\pm 17.1\%$ ) | 97.5%<br>( $\pm 5\%$ )   | 100%<br>( $\pm 0\%$ )    | 87.9%          |
| <b>Kidd44<br/>(7 Pops)</b>   | 57.5%<br>( $\pm 9.6\%$ ) | 85%<br>( $\pm 12.9\%$ )   | 55%<br>( $\pm 5.8\%$ )    | 72.5%<br>( $\pm 12.6\%$ ) | 55%<br>( $\pm 12.9\%$ )   | 92.5%<br>( $\pm 9.6\%$ ) | 82.5%<br>( $\pm 9.6\%$ ) | 71.4%          |

**Table 4: MLR Classification Accuracy**

Comparison of the nine possible combinations of each of three simulated datasets on each of three SNP panels and their MLR classification accuracy for each population. Reported as percent accuracy with two-tailed standard deviations listed in parentheses (). Abbreviations used: GOAL = Genomic Origins and Admixture in Latinos, 1000G = 1000 Genomes Project, 7 Pops = 7 Populations Combined, COL = Colombia, CUB = Cuba, DOM = Dominican Republic, HUR = Honduras, PUR = Puerto Rico, PEL = Peru from Lima, MXL = Mexicans living in Los Angeles, and MLR = multinomial logistic regression. Both Colombian populations from GOAL and 1000G are listed in this table as “COL”.

## CHAPTER 4

### *Conclusions and Future Directions*

Cassandra Hernandez Setser

John V. Planz

Robert C. Barber

Nicole R. Phillips

Ranajit Chakraborty

Deanna S. Cross

## 4.1 Summary

The Hispanic BGA AIMs (ancestry informative markers) panels described here, Setser234 and Setser80, are able to differentiate biogeographic ancestry (BGA) of individuals of Hispanic origin, particularly those surrounding the Caribbean. Based on the Genomic Origins and Admixture in Latinos (GOAL) dataset<sup>1</sup>, I designed my panels using country attributable mean  $F_{ST}$ , a variant of Wright's  $F_{ST}$ <sup>2</sup>, for SNP ascertainment focused on distinguishing countries through dedicated SNPs. Utilizing this method I was able to select 234 SNPs from a filtered dataset of 494,886 SNPs, and further compress my panel by 65.8% by adjusting the proportion of the SNPs attributed to Honduras (HUR) and the Dominican Republic (DOM) in favor of those attributed to Colombia (COL), Cuba (CUB), and Puerto Rico (PUR). I found that these 80 SNPs were sufficient for differentiation to 95% accuracy in the GOAL dataset and it performed better than the Kidd panel<sup>3</sup> and comparably if not better than the Seldin panel<sup>4</sup> on an expanded dataset. When combining the GOAL dataset with the Admixed American dataset from the 1000 Genomes Project<sup>5</sup>, the Setser80 performed to >90% accuracy overall, including two populations not involved in SNP ascertainment, Mexicans living in Los Angeles (MXL, >90%) and Peru in Lima (PEL, >98%).

While panel selection was successful, there are a number of potential limitations to my study. These include SNP selection and a lack of inclusion of neighboring countries within my testing groups. Despite these limitations there are a number of fields of study that my panel could contribute to now and in the future including forensics, genealogy, and potentially precision medicine.

## 4.2 Future Directions

### 4.2.1 Potential Improvements

As with all research, there are various improvements that could be made. To balance the panel, 77.4% of the SNPs attributed to Honduras were removed; this unintentionally also removed a disproportionate amount of SNPs attributed to Puerto Rico. Inclusion of additional Puerto Rico SNPs would be beneficial, particularly those with high PUR vs. CUB  $F_{ST}$ , given the difficulty separating these two populations by STRUCTURE<sup>6</sup> and PCA. Query of SNP databases such as AncestrySNPminer<sup>7</sup> could produce additional SNPs of value.

### 4.2.2 Sample Size

The size of the unrelated individuals GOAL dataset ( $n=160$ )<sup>1</sup> is a limiting factor in this study. I combined this dataset with the Admixed American super-population from the 1000 Genomes Project Phase III ( $n=347$ )<sup>5</sup> to address this, but other studies have much larger populations. It would be interesting to test this panel on a mega-dataset by assembling data from multiple studies such as: POPRES ( $n=205$  including Mexicans from Guadalajara)<sup>8</sup>, Seldin128 ( $n=825$  including Mexican, Puerto Rican, and Amerindian)<sup>4</sup>, Human Genome Diversity Project – Centre d'Étude du Polymorphisme Humain (HGDP-CEPH;  $n=938$  including Colombia, Pima, Maya, Surui, Karitiana, and Yoruba)<sup>9</sup>, the Diversity of Latin Americans study ( $n=7342$  from Brazil, Chile, Colombia, Mexico, and Peru)<sup>10</sup>, and the Hispanic Community Health Survey/Study of Latinos (HCHS/SOL;  $n=12,803$  from Cuba, Dominican Republic, Puerto Rican, Mexican, and Central and South American)<sup>11</sup>. The Genome Aggregation Database (gnomAD) has made great strides towards dataset aggregation and gnomAD v3.0 contains a total of 71,702 samples, 6,835 of which are Latino/Admixed American<sup>12</sup>. The reference population on which an AIMs panel was designed has a major impact of the accuracy of the panel.

Testing my panel's ability to separate populations from expanded datasets with new populations and further adjusting it would greatly improve genetic differentiation for Central and South American regions. Sequence data, including mitochondrial and Y chromosome, would be invaluable because the GOAL study<sup>1</sup> used an older generation genechip<sup>13</sup> designed using HapMap I data<sup>14</sup>, which may not capture the full diversity of Hispanic populations.

#### **4.2.3 Geography**

The labels provided in the datasets, which corresponded to country and not direct geographical space, also confined my Hispanic BGA classification. Natural topographical borders (e.g. mountains and bodies of water) are far more relevant than arbitrary geopolitical borders (e.g. country) for ancestry determination. For example, the three Western Antilles populations (CUB, PUR, and DOM) are separated from HUR by the Caribbean Sea and were easily distinguishable. What remains to be seen is if the Setser80 AIMs, which differentiated HUR in this study, retain sufficient  $F_{ST}$  to distinguish Honduras from the two adjacent populations of the Northern Triangle: Guatemala and El Salvador. PCA analysis conducted in Chapter 2 had CUB clustering in one section that tailed off along PC1 and overlapped with DOM, similar to that seen in the Hispanic Community Health Survey (HCHS)<sup>11</sup>. Contrastingly, Colombia would seem to be isolated from the Western Antilles; however, mtDNA and linguistic evidence suggests the Caribbean was populated from the Orinoco Basin<sup>15-17</sup>. In this study, it should be noted that none of the seven populations share a geographic border.

### **4.3 Applications**

Even with these potential limitations or areas of improvement, my panel could be utilized for forensic genetics, genealogical genetics, and potentially clinical genetics; all of these fields use allele frequencies to answer various questions.

#### **4.3.1 Forensic Genetics**

The Setser80 panel has the potential to be utilized to help identify unidentified human remains. There are over 40,000 recalcitrant UHRs in the US at any time, a fact that has been referred to as a “mass disaster over time”<sup>18</sup>. According to 2010 Census data, Hispanic individuals are 17% of the total US population (and the majority of the minority groups) and, as of 2012, became the largest population group of those under age 18, and are projected to be 38% of the population of the United States by 2060<sup>19</sup>. The Hispanic demographic continues to grow in our nation of immigrants where 60.1% of the 5,749,343 individuals who entered the country between 2005 and 2010 originated from Central and South America<sup>20</sup>. These individuals have origins across two continents and, despite their heterogeneity, are all grouped under the catchall terms “Hispanic” or “Latino”. In fact, the allele frequencies used to calculate random match probability in forensic genetics were originally divided into Southeast Hispanics (Puerto Rico, Cuba, etc.) and Southwest Hispanics (Mexico), leaving the choice of which population’s allele frequencies to use at the lab’s discretion or selected on a case by case basis pursuant to details of the case itself<sup>21</sup>. Hispanic UHRs found on the US-Mexico border are often assumed to be of Southwest Hispanic origin, but if that assumption is inaccurate it can have a considerable effect on the reported conclusions. The Setser80 panel could create a more distinct classification by creating a likely country of origin classification.

Ancestry informative markers (AIMs) are used for objective human identification of deceased individuals, both modern and ancient. The Setser80 panel could be utilized to determine country of origin in mass graves or mass disaster areas within Central and South America. The small size of this panel lends itself to multiplexing to conserve precious specimen for future analysis.

#### **4.3.2 Genetic Genealogy**

In the past decade the cost of DNA analysis has become affordable to individual consumers, which has prompted an entire cottage industry that has grown up around direct-to-consumer (DTC) ancestry DNA kits. One of the stated goals of individuals who participate in these tests is to find out their country of origin. The Setser80 panel I developed or similar panels could be used to help stratify individuals into a country because there is still wide variability in the accuracy of these tests<sup>22, 23</sup>. The DTC DNA kits came to market with Rite Aid's HomeDNA Paternity Test kits in 2007<sup>24</sup>. Since then the market has matured, and now two main companies that account for most of the DTC ancestry DNA analysis in the United States: AncestryDNA and 23andMe.

AncestryDNA created their reference database using 800 publicly available references from HGDP-CEPH, 1,500 proprietary in-house reference samples, and 1,800+ samples from their customers<sup>25</sup>. Those ~4,100 references were trimmed by removing related samples based on identity-by-descent and outlier samples based on PCA and was further refined using leave-one-out cross-validation<sup>25</sup>. Of these, only 281 samples (9.4% of the database) were from Hispanic relevant populations: 83 references from Iberia, 131 Native Americans, and 67 from Nigeria<sup>25</sup>. The actual classification of customer ethnicity is based on the algorithm applied in ADMIXTURE<sup>25, 26</sup>, a data analysis program very similar to STRUCTURE<sup>6</sup>. My panel could add

additional granularity to the country of origin results by highlighting polymorphisms that could be of high value for separating the Hispanic population.

Similarly, 23andMe performs their analysis as a five-step process: phasing, window classification (using a support vector machine), smoothing, re-calibrating (using simulated data), and aggregation & reporting<sup>27</sup>. As of 2014, 23andMe uses a database of 10,699 individuals from 45 populations including Balkan, Iberian, Italian, and Sardinian from Southern Europe; Senegambian & Guinean, Coastal West Africa, and Nigerian from West Africa; and a conglomerate Native American population under East Asian & Native American<sup>27</sup>. Similar to AncestryDNA, and in the context of Hispanic populations, my panel could be utilized to highlight potential high value polymorphisms for distinguishing these populations. In a broader sense some of my methodology such as country attributable  $F_{ST}$  might identify new polymorphisms that could be used for any closely related populations.

### **4.3.3 Clinical Genetics**

Complex admixture means Hispanics have not been studied as extensively as European ancestry. A simple query of PubMed for “GWAS” AND population (“Hispanic”, “Latino”, “African American”, “European”, or “Caucasian”) yields the following: African American = 1053, Caucasian = 2,586, European = 5,145, Hispanic = 547, and Latino = 373 as of January 6, 2020. Acknowledging the likely overlap between “Hispanic” or Latino” and between “European” or “Caucasian”, I conservatively estimated complete overlap and considered only the label with the most citations per pair where Hispanic and Latino ( $n = 547$ ) *still* have far fewer GWAS studies than European and Caucasian ( $n = 5,145$ ) despite efforts to combat health disparities. While GWAS studies between ancestry and disease status exist, typical association studies avoid consideration of the difference in allele frequencies between the populations. That

is, population substructure is not corrected for automatically. My Setser80 panel could potentially be used to help correct for substructure within this highly admixed population.

It is not known how similar the conglomerate Hispanic demographic in the United States (predominantly Mexican) is to Central American populations from the Northern Triangle (Guatemala, Honduras, and El Salvador); therefore, it is important to characterize their genetics to avoid assuming their disease risk or drug metabolism operates in the same way as other Hispanic populations<sup>28</sup>. My Hispanic AIMs panel, in combination with full disease screening of individuals from the Northern Triangle, could be used for inexpensive and rapid determination of ancestry that will allow practitioners to focus primarily on the diseases of the individuals' specific ancestry. Studies demonstrate that Hispanic individuals have differential predispositions for various diseases, while acknowledging the diversity among Latin American populations restricts the results only to the populations studied<sup>29</sup>. As early as 2003, it was known that Southwestern Hispanics (Mexico) were predisposed to obesity<sup>30</sup>, diabetes<sup>31</sup>, and gallbladder<sup>32</sup> disease while Southeastern Hispanics (Puerto Rico) were predisposed to hypertension<sup>33,21</sup>. Both of these populations are Hispanic, but considering them as a single population could have drastic effects in their medical care.

#### **4.4 Ethical Considerations**

Ethical and legal considerations must also be accounted for in the study of ancestry informative markers. The forensic genetic community and the public are currently struggling with how best to regulate the utilization of genetic information for identification and classification purposes. Using public genetic genealogy databases such as GEDMatch<sup>34</sup> for criminal investigations has only recently been regulated<sup>35</sup>, because it is the relative that has consented to the use of their genetic data, not the suspect themselves as in the FBI's Combined

DNA Index System (CODIS) Arrestee and Convicted Offender databases<sup>36</sup>. In the United States, The Genetic Information and Nondiscrimination Act (GINA) only covers discrimination based on genetics as it applies to employment and health care<sup>37</sup>. However, AIMs genetic data can also be used for discrimination in other ways, such as the cataloguing of ethnic minority Uyghurs in China<sup>38, 39</sup>. We must take care to preserve the anonymity of genetic data (as much as is possible) and educate lawmakers on what is appropriate scientific use of such data.

Despite these ethical considerations, utilizing an identification panel can also have great benefits. My exploration of the human genome for AIMs that can differentiate Hispanic populations was originally inspired by the influx of unaccompanied minors in the summer of 2014. From 2011 to 2014, there was a steady increase of unaccompanied minors from Central and South America<sup>40</sup>. Hispanic ancestry studies continue to be relevant given the continued influx of people seeking to cross the US-Mexico border. One of the arguments for the 2018 family separation policy was an effort to combat human trafficking, the rationale being that children may have been abducted to be presented at the border as part of a family and increase the chances of the adult(s) gaining entry and claiming asylum. While normal autosomal paternity (or maternity) DNA analysis can inform immigration authorities that the adult is or is not the parent or close relation, ancestry DNA can be used to identify where the child came from and what country/region to look for missing persons reports.

## 4.5 References

- 1) Moreno-Estrada, A. *et al.* Reconstructing the population genetic history of the Caribbean. *PLoS Genet.* **9** (11), e1003925; 10.1371/journal.pgen.1003925 (2013 November 14).
- 2) Wright, S. The genetical structure of populations. *Ann Eugen.* **15** (4), 323-354 (1951 March).

- 3) Kidd, K. K., Kidd, J. R., Pakstis, A. J., Speed, W. C. & Donnelly, M. P. Developing SNP Panels for ancestry identification useful in forensic investigations. Poster. (2011).
- 4) Kosoy, R. *et al.* Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat.* **30** (1), 69-78 (2009 January).
- 5) Auton, A. *et al.* A global reference for human genetic variation. *Nature.* **526** (7571), 68-74 (2015 September 30).
- 6) Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics.* **155** (2), 945-959 (2000 June).
- 7) Amirisetty, S., Hershey, G. K. K. & Baye, T. M. AncestrySNPminer: A bioinformatics tool to retrieve and develop ancestry informative SNP panels. *Genomics.* **100** (1), 57-63 (2012 July).
- 8) Nelson, M. R. *et al.* The Population Reference Sample, POPRES: A Resource for population, disease, and pharmacological genetics research. *Am J Hum Genet.* **83** (3), 347-358 (2008 September 12).
- 9) Cann, H. M. *et al.* A human genome diversity cell line panel. *Science.* **296** (5566), 261-262 (2002 April 12).
- 10) Ruiz-Linares, A. *et al.* Admixture in Latin America: Geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet.* **10** (9), e1004572; 10.1371/journal.pgen.1004572 (2014 September 1).
- 11) Conomos, M. P. *et al.* Genetic diversity and association studies in US Hispanic/Latino populations: Applications in the Hispanic Community Health Study/Study of Latinos. *Am J Hum Genet.* **98** (1), 165-184 (2016 January 7).
- 12) Francioli, L. & MacArthur, D. gnomAD v3.0. MacArthur Lab Blog. (2019 October 16). URL <https://macarthurlab.org/blog/>
- 13) McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet.* **40** (10), 1166-1174 (2008 October).
- 14) The International HapMap Consortium. The International HapMap Project. *Nature.* **426** (6968), 789-796 (2003 December 18).
- 15) Lalueza-Fox, C., Gilbert, M. T. P., Martínez-Fuentes, A. J., Calafell, F. & Bertranpetit, J. Mitochondrial DNA from Pre-Columbian Ciboneys from Cuba and the prehistoric colonization of the Caribbean. *Am J of Phys Anthro.* **121** (2), 97-108 (2003 June 1).

- 16) Bodner, M. *et al.* Rapid coastal spread of First Americans: Novel insights from South America's Southern Cone mitochondrial genomes. *Genome Res.* **22** (5), 811-820 (2012 May).
- 17) Marcheco-Teruel, B. *et al.* Cuba: Exploring the history of admixture and the genetic basis of pigmentation using autosomal and uniparental markers. *PLoS Genet.* **10** (7), e1004488; 10.1371/journal.pgen.1004488 (2014 July 24).
- 18) Ritter, N. Missing persons and unidentified remains: The nation's silent mass disaster. *NIJ Journal.* **256** (2007).
- 19) Frey, W. H. Census projects new "majority minority" tipping points. *Brookings.* (2012 December 13). URL <https://www.brookings.edu/opinions/census-projects-new-majority-minority-tipping-points/>
- 20) Abel, G. J. & Sander, N. Quantifying global international migration flows. *Science.* **343** (6178), 1520-1522 (2014 March 28).
- 21) Bertoni, B., Budowle, B., Sans, M., Barton, S.A. & Chakraborty, R. Admixture in Hispanics: Distribution of ancestral population contributions in the continental United States. *Hum Bio.* **75** (1), 1-11 (2003 February).
- 22) Saey, T. H. DNA testing can bring families together, but gives mixed answers on ethnicity. *ScienceNews.* (2018 June 23).
- 23) Tandy-Connor, S. *et al.* False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care. *Genet in Med.* **20** (12), 1515-1521 (2018 December 1).
- 24) Rite Aid HomeDNA Paternity Test, 2007. URL <https://www.riteaid.com/shop/homedna-paternity-test-for-at-home-use-1-ct-8022392>.
- 25) Ball, C. *et al.* Ethnicity Estimate White Paper. *ancestryDNA.* (2013 October 30).
- 26) Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19** (9), 1655-1664 (2009 September).
- 27) Durand, E. Y., Do, C. B., Mountain, J. L. & Macpherson, J. M. Ancestry composition: A novel, efficient pipeline for ancestry deconvolution. 23andMe White Paper 23-16. (2014 October 17).
- 28) Tishkoff, S. A. & Kidd, K. K. Implications of biogeography of human populations for 'race' and medicine. *Nature Genet.* **36** (11S), S21-S27; 10.1038/ng1438 (2004 November).

- 29) Manichaikul, A. *et al.* Population structure of Hispanics in the United States: The multi-ethnic study of atherosclerosis. *PLoS Genet.* **8** (4), e1002640; 10.1371/journal.pgen.1002640 (2012 April 12).
- 30) Gao, C., *et al.* A comprehensive analysis of common and rare variants to identify adiposity loci in Hispanic Americans: The IRAS Family Study (IRASFS). *PLoS ONE.* **10** (11), e0134649; 10.1371/journal.pone.0134649 (2015 November 1).
- 31) Granados-Silvestre, M. A. *et al.* Susceptibility background for type 2 diabetes in eleven Mexican Indigenous populations: HNF4A gene analysis. *Mol Genet Genomics.* **292** (6), 1209-1219 (2017 December 1).
- 32) Price, A. L. *et al.* A genomewide admixture map for Latino populations. *Am J Hum Genet.* **80** (6), 1024-1036 (2007 June).
- 33) Pabon-Nau, L. P., Cohen, A., Meigs, J. B. & Grant, R. W. Hypertension and diabetes prevalence among U.S. Hispanics by country of origin: The National Health Interview Survey 2000-2005. *J Gen Intern Med.* **25** (8), 847-852 (2010 August).
- 34) Russell, J. G. Gedmatch: a DNA geek's dream site. *The Legal Genealogist.* (2012 August 12). URL <https://www.legalgenealogist.com/2012/08/12/gedmatch-a-dna-geeks-dream-site/>
- 35) Russell, J. G. GEDmatch reverses course. *The Legal Genealogist.* (2019 May 19). URL <https://www.legalgenealogist.com/2019/05/19/gedmatch-reverses-course/>
- 36) National DNA Index System (NDIS) Operations Manual, version 8. FBI Laboratory. (2019 May 1).
- 37) Genetic Information Nondiscrimination Act of 2008, Pub. L. No. 110-233 (2008). <https://www.eeoc.gov/laws/statutes/gina.cfm>
- 38) Wee, S. China uses DNA to track its people, with the help of American expertise. *New York Times.* (2019 February 21).
- 39) Jin, X.-Y. *et al.* A set of novel SNP loci for differentiating continental populations and three Chinese populations. *PeerJ.* **7**, e6508; 10.7717/peerj.6508 (2019 March 29).
- 40) Negroponte, D. V. The surge in unaccompanied children from Central America: A humanitarian crisis at our border. *Brookings.* (2014 July 2). URL <https://www.brookings.edu/blog/up-front/2014/07/02/the-surge-in-unacc...d-children-from-central-america-a-humanitarian-crisis-at-our-border/>

## APPENDIX

**Supplemental Table S2.1:  $F_{ST}$  Statistic for the Setser80**

| SNP        | COL vs.<br>CUB<br>$F_{ST}$ | COL vs.<br>DOM<br>$F_{ST}$ | COL vs.<br>HUR<br>$F_{ST}$ | COL vs.<br>PUR<br>$F_{ST}$ | CUB vs.<br>DOM<br>$F_{ST}$ | CUB vs.<br>HUR<br>$F_{ST}$ | CUB vs.<br>PUR<br>$F_{ST}$ | DOM vs.<br>HUR<br>$F_{ST}$ | DOM vs.<br>PUR<br>$F_{ST}$ | HUR vs.<br>PUR<br>$F_{ST}$ | Mean<br>$F_{ST}$ |
|------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|------------------|
| rs12130873 | 0.02136                    | 0.01523                    | 0.23743                    | 0.01372                    | 0.00264                    | 0.23168                    | 0.00118                    | 0.31409                    | 0.00003*                   | 0.33016                    | 0.11675          |
| rs6694897  | 0.14630                    | 0.15723                    | 0.18921                    | 0.16943                    | 0.14101                    | 0.08312                    | 0.08346                    | 0.03199                    | 0.02449                    | 0.01305*                   | 0.10132          |
| rs1570099  | 0.12924                    | 0.14424                    | 0.16141                    | 0.16369                    | 0.12662                    | 0.06736                    | 0.06373                    | 0.03474                    | 0.03218                    | 0.01291*                   | 0.09103          |
| rs10493701 | 0.04411                    | 0.00876                    | 0.20767                    | 0.00724                    | 0.07290                    | 0.18628                    | 0.06543                    | 0.22719                    | 0.00187*                   | 0.23910                    | 0.10568          |
| rs1040424  | 0.06061                    | 0.00125                    | 0.14142                    | 0.00294                    | 0.08222                    | 0.15247                    | 0.07259                    | 0.15904                    | 0.01465*                   | 0.16933                    | 0.08272          |
| rs10495889 | 0.05877                    | 0.03008                    | 0.16038                    | 0.03825                    | 0.11837                    | 0.14681                    | 0.06727                    | 0.16385                    | 0.02373                    | 0.17784                    | 0.09853          |
| rs17018313 | 0.01280                    | 0.00668*                   | 0.19987                    | 0.00805*                   | 0.01894                    | 0.19414                    | 0.01711                    | 0.26947                    | 0.00256*                   | 0.28424                    | 0.09793          |
| rs7568419  | 0.04844                    | 0.02191                    | 0.16717                    | 0.02062                    | 0.04850                    | 0.17164                    | 0.03463                    | 0.24218                    | 0.04174                    | 0.30209                    | 0.10989          |
| rs11693873 | 0.06161                    | 0.03212                    | 0.17206                    | 0.03240                    | 0.08637                    | 0.14060                    | 0.08297                    | 0.14644                    | 0.00038*                   | 0.15187                    | 0.09061          |
| rs4433950  | 0.01326                    | 0.00651*                   | 0.20323                    | 0.00790*                   | 0.03187                    | 0.19048                    | 0.02938                    | 0.26195                    | 0.00345                    | 0.27627                    | 0.09955          |
| rs9853146  | 0.14228                    | 0.12291                    | 0.18906                    | 0.12965                    | 0.15898                    | 0.13224                    | 0.13075                    | 0.04184                    | 0.00311                    | 0.02862                    | 0.10794          |
| rs4685443  | 0.11960                    | 0.14048                    | 0.14544                    | 0.15009                    | 0.08436                    | 0.04481                    | 0.03160                    | 0.04611                    | 0.03052                    | 0.01054*                   | 0.07825          |
| rs259425   | 0.13624                    | 0.16602                    | 0.18872                    | 0.15814                    | 0.10776                    | 0.03165                    | 0.04636                    | 0.06488                    | 0.07145                    | 0.00325*                   | 0.09680          |
| rs2356298  | 0.07992                    | 0.04926                    | 0.19876                    | 0.04926                    | 0.11418                    | 0.15765                    | 0.09575                    | 0.15429                    | 0.00334                    | 0.15792                    | 0.10603          |
| rs11719358 | 0.09912                    | 0.13420                    | 0.16491                    | 0.12028                    | 0.15724                    | 0.06468                    | 0.05171                    | 0.11445                    | 0.08203                    | 0.02709                    | 0.10157          |
| rs9857908  | 0.12265                    | 0.15319                    | 0.21601                    | 0.15047                    | 0.20331                    | 0.10969                    | 0.07975                    | 0.15533                    | 0.09551                    | 0.07984                    | 0.13657          |
| rs1586861  | 0.16504                    | 0.15315                    | 0.19918                    | 0.16167                    | 0.14848                    | 0.13311                    | 0.13746                    | 0.00633                    | 0.01097*                   | 0.00315                    | 0.10966          |
| rs3910480  | 0.12445                    | 0.14086                    | 0.15129                    | 0.14969                    | 0.05526                    | 0.04296                    | 0.07184                    | 0.00175                    | 0.03367                    | 0.02424                    | 0.07960          |
| rs1366363  | 0.07582                    | 0.01519                    | 0.19465                    | 0.02053                    | 0.11305                    | 0.19362                    | 0.08934                    | 0.20508                    | 0.00541*                   | 0.22018                    | 0.11220          |
| rs3733838  | 0.04389                    | 0.00248                    | 0.19398                    | 0.00446                    | 0.06036                    | 0.18333                    | 0.06874                    | 0.22558                    | 0.00476*                   | 0.24026                    | 0.10183          |
| rs1438745  | 0.02533                    | 0.18021                    | 0.00563                    | 0.05249                    | 0.18034                    | 0.00320*                   | 0.04119                    | 0.26755                    | 0.26260                    | 0.04352                    | 0.10557          |
| rs16902270 | 0.14557                    | 0.16233                    | 0.18446                    | 0.16684                    | 0.11799                    | 0.06752                    | 0.07443                    | 0.02803                    | 0.02933                    | 0.01502*                   | 0.09615          |
| rs871234   | 0.07459                    | 0.17146                    | 0.09206                    | 0.05378                    | 0.15663                    | 0.01475                    | 0.00150                    | 0.24140                    | 0.24402                    | 0.03115                    | 0.10813          |
| rs692713   | 0.09401                    | 0.18700                    | 0.11996                    | 0.07294                    | 0.16283                    | 0.01292                    | 0.01075*                   | 0.24732                    | 0.22844                    | 0.01679                    | 0.11315          |
| rs4608884  | 0.12067                    | 0.13477                    | 0.13419                    | 0.17569                    | 0.15144                    | 0.08472                    | 0.07586                    | 0.05550                    | 0.06161                    | 0.01129                    | 0.10057          |
| rs190592   | 0.11932                    | 0.19870                    | 0.12133                    | 0.08898                    | 0.16337                    | 0.00291                    | 0.00792*                   | 0.24295                    | 0.23473                    | 0.00081                    | 0.11652          |
| rs6596807  | 0.04105                    | 0.02010                    | 0.20048                    | 0.01865                    | 0.07882                    | 0.16821                    | 0.07273                    | 0.21379                    | 0.02509                    | 0.22430                    | 0.10632          |
| rs9392285  | 0.07138                    | 0.01935                    | 0.19268                    | 0.02448                    | 0.08596                    | 0.18517                    | 0.12283                    | 0.19751                    | 0.00728                    | 0.21554                    | 0.11222          |

| SNP        | COL vs.<br>CUB<br>F <sub>ST</sub> | COL vs.<br>DOM<br>F <sub>ST</sub> | COL vs.<br>HUR<br>F <sub>ST</sub> | COL vs.<br>PUR<br>F <sub>ST</sub> | CUB vs.<br>DOM<br>F <sub>ST</sub> | CUB vs.<br>HUR<br>F <sub>ST</sub> | CUB vs.<br>PUR<br>F <sub>ST</sub> | DOM vs.<br>HUR<br>F <sub>ST</sub> | DOM vs.<br>PUR<br>F <sub>ST</sub> | HUR vs.<br>PUR<br>F <sub>ST</sub> | Mean<br>F <sub>ST</sub> |
|------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-------------------------|
| rs1329521  | 0.15709                           | 0.15311                           | 0.18385                           | 0.17877                           | 0.15953                           | 0.11346                           | 0.10928                           | 0.01519                           | 0.01231                           | 0.00961*                          | 0.10730                 |
| rs17745021 | 0.00602*                          | 0.00726                           | 0.18035                           | 0.01559                           | 0.00890                           | 0.17775                           | 0.01727                           | 0.25347                           | 0.04501                           | 0.27456                           | 0.09741                 |
| rs3777908  | 0.01129                           | 0.00482                           | 0.19754                           | 0.00327                           | 0.00583                           | 0.19560                           | 0.00425                           | 0.27318                           | 0.00207                           | 0.28808                           | 0.09859                 |
| rs17087570 | 0.06604                           | 0.00537*                          | 0.22108                           | 0.00546*                          | 0.07436                           | 0.22732                           | 0.08779                           | 0.27261                           | 0.01493*                          | 0.28288                           | 0.12063                 |
| rs4709836  | 0.11661                           | 0.15187                           | 0.14770                           | 0.16086                           | 0.12933                           | 0.03893                           | 0.03556                           | 0.08065                           | 0.07383                           | 0.01132*                          | 0.09240                 |
| rs12536738 | 0.07829                           | 0.03401                           | 0.20886                           | 0.03500                           | 0.10623                           | 0.17465                           | 0.09867                           | 0.17456                           | 0.00853*                          | 0.18462                           | 0.10863                 |
| rs10953750 | 0.06575                           | 0.01883                           | 0.16149                           | 0.01933                           | 0.08834                           | 0.15363                           | 0.08909                           | 0.15051                           | 0.00990*                          | 0.16034                           | 0.08974                 |
| rs2352479  | 0.09944                           | 0.12592                           | 0.12388                           | 0.16367                           | 0.12004                           | 0.04338                           | 0.04705                           | 0.05936                           | 0.07068                           | 0.01437                           | 0.08678                 |
| rs17480133 | 0.17271                           | 0.19398                           | 0.22183                           | 0.19761                           | 0.12585                           | 0.06882                           | 0.08521                           | 0.03356                           | 0.04229                           | 0.00784*                          | 0.11340                 |
| rs766382   | 0.10783                           | 0.09639                           | 0.14531                           | 0.10471                           | 0.15439                           | 0.10352                           | 0.09324                           | 0.05682                           | 0.02488                           | 0.02515                           | 0.09122                 |
| rs1588459  | 0.16094                           | 0.17730                           | 0.18981                           | 0.19184                           | 0.10978                           | 0.08773                           | 0.07225                           | 0.02378                           | 0.01105                           | 0.01087*                          | 0.10136                 |
| rs6474712  | 0.06351                           | 0.20209                           | 0.04217                           | 0.03827                           | 0.19691                           | 0.00109                           | 0.00100*                          | 0.28843                           | 0.26641                           | 0.01229*                          | 0.10856                 |
| rs880397   | 0.08331                           | 0.18444                           | 0.02767                           | 0.03029                           | 0.17970                           | 0.01329                           | 0.01599                           | 0.27002                           | 0.24699                           | 0.01705*                          | 0.10346                 |
| rs10981894 | 0.09794                           | 0.07058                           | 0.10767                           | 0.13719                           | 0.18782                           | 0.11750                           | 0.12768                           | 0.07918                           | 0.10264                           | 0.13199                           | 0.11602                 |
| rs2008617  | 0.04151                           | 0.23629                           | 0.03781                           | 0.05985                           | 0.24319                           | 0.00878*                          | 0.00819                           | 0.33617                           | 0.31925                           | 0.01496                           | 0.12884                 |
| rs16912280 | 0.06727                           | 0.22852                           | 0.06626                           | 0.04629                           | 0.22431                           | 0.00910                           | 0.00491*                          | 0.32309                           | 0.29758                           | 0.00340                           | 0.12609                 |
| rs1259603  | 0.02251                           | 0.01626                           | 0.22770                           | 0.02316                           | 0.05286                           | 0.18333                           | 0.06271                           | 0.25254                           | 0.04393                           | 0.27361                           | 0.11586                 |
| rs16932385 | 0.02925                           | 0.01002                           | 0.26014                           | 0.00765                           | 0.06678                           | 0.22826                           | 0.06055                           | 0.30385                           | 0.02558                           | 0.31814                           | 0.13102                 |
| rs11189628 | 0.14119                           | 0.15050                           | 0.18026                           | 0.15521                           | 0.10708                           | 0.07625                           | 0.09228                           | 0.00541                           | 0.01201                           | 0.00759*                          | 0.09126                 |
| rs17112705 | 0.15421                           | 0.17786                           | 0.20491                           | 0.18132                           | 0.14930                           | 0.07153                           | 0.07455                           | 0.06097                           | 0.05382                           | 0.00910*                          | 0.11194                 |
| rs1849352  | 0.04784                           | 0.02982                           | 0.18714                           | 0.02097                           | 0.09955                           | 0.15449                           | 0.05584                           | 0.19384                           | 0.01604                           | 0.19270                           | 0.09982                 |
| rs2878712  | 0.02350                           | 0.01319                           | 0.21765                           | 0.02009                           | 0.05090                           | 0.17748                           | 0.06149                           | 0.24141                           | 0.03553                           | 0.26197                           | 0.11032                 |
| rs10840730 | 0.03540                           | 0.00242*                          | 0.18961                           | 0.00952*                          | 0.05926                           | 0.18939                           | 0.03252                           | 0.25024                           | 0.00123*                          | 0.27258                           | 0.10158                 |
| rs2051827  | 0.13258                           | 0.14135                           | 0.21359                           | 0.14807                           | 0.09160                           | 0.09166                           | 0.14127                           | 0.03493                           | 0.03900                           | 0.07688                           | 0.11109                 |
| rs12146822 | 0.12337                           | 0.06844                           | 0.22311                           | 0.06976                           | 0.14798                           | 0.19816                           | 0.14918                           | 0.14798                           | 0.01263*                          | 0.15767                           | 0.12730                 |
| rs7310083  | 0.09213                           | 0.10322                           | 0.11711                           | 0.21128                           | 0.00655*                          | 0.00630*                          | 0.19156                           | 0.01749*                          | 0.22097                           | 0.25073                           | 0.11567                 |
| rs1967232  | 0.13608                           | 0.14002                           | 0.16603                           | 0.14445                           | 0.11483                           | 0.08604                           | 0.09842                           | 0.00013*                          | 0.00396                           | 0.01392*                          | 0.08758                 |
| rs6486527  | 0.09037                           | 0.11347                           | 0.10785                           | 0.16676                           | 0.13478                           | 0.05597                           | 0.06954                           | 0.05911                           | 0.09296                           | 0.04229                           | 0.09331                 |
| rs9569702  | 0.05638                           | 0.01380                           | 0.15060                           | 0.01385                           | 0.07090                           | 0.14390                           | 0.08272                           | 0.14845                           | 0.00957*                          | 0.15819                           | 0.08292                 |
| rs4341647  | 0.05664                           | 0.01630                           | 0.16065                           | 0.01417                           | 0.06451                           | 0.15779                           | 0.10149                           | 0.18030                           | 0.00407                           | 0.18691                           | 0.09428                 |

| SNP               | COL vs.<br>CUB<br>F <sub>ST</sub> | COL vs.<br>DOM<br>F <sub>ST</sub> | COL vs.<br>HUR<br>F <sub>ST</sub> | COL vs.<br>PUR<br>F <sub>ST</sub> | CUB vs.<br>DOM<br>F <sub>ST</sub> | CUB vs.<br>HUR<br>F <sub>ST</sub> | CUB vs.<br>PUR<br>F <sub>ST</sub> | DOM vs.<br>HUR<br>F <sub>ST</sub> | DOM vs.<br>PUR<br>F <sub>ST</sub> | HUR vs.<br>PUR<br>F <sub>ST</sub> | Mean<br>F <sub>ST</sub> |
|-------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-------------------------|
| <b>rs9556940</b>  | 0.14825                           | 0.16077                           | 0.17331                           | 0.16619                           | 0.09477                           | 0.08631                           | 0.08767                           | 0.00633*                          | 0.00356*                          | 0.01316*                          | 0.08942                 |
| <b>rs1957572</b>  | 0.05095                           | 0.03394                           | 0.17552                           | 0.02228                           | 0.11559                           | 0.15235                           | 0.05842                           | 0.19330                           | 0.02768                           | 0.18830                           | 0.10183                 |
| <b>rs178384</b>   | 0.10567                           | 0.07545                           | 0.13410                           | 0.10048                           | 0.15544                           | 0.11984                           | 0.10358                           | 0.05628                           | 0.02126                           | 0.05578                           | 0.09279                 |
| <b>rs12434466</b> | 0.06045                           | 0.00397                           | 0.18396                           | 0.00635                           | 0.08965                           | 0.18450                           | 0.07129                           | 0.20759                           | 0.00935*                          | 0.22593                           | 0.10243                 |
| <b>rs17094860</b> | 0.05559                           | 0.01398                           | 0.21462                           | 0.00702                           | 0.10716                           | 0.20471                           | 0.06533                           | 0.25468                           | 0.00927                           | 0.25856                           | 0.11909                 |
| <b>rs12435621</b> | 0.01696                           | 0.00164*                          | 0.27281                           | 0.00235*                          | 0.01966                           | 0.26739                           | 0.01881                           | 0.35841                           | 0.00348                           | 0.37576                           | 0.13293                 |
| <b>rs12431505</b> | 0.03591                           | 0.00398*                          | 0.34602                           | 0.00119*                          | 0.05134                           | 0.33386                           | 0.05551                           | 0.43078                           | 0.01278                           | 0.45007                           | 0.17111                 |
| <b>rs1462266</b>  | 0.05430                           | 0.00750*                          | 0.21034                           | 0.00693*                          | 0.07260                           | 0.21682                           | 0.05326                           | 0.26940                           | 0.00805*                          | 0.29686                           | 0.11511                 |
| <b>rs17097005</b> | 0.01992                           | 0.00340*                          | 0.19965                           | 0.00324                           | 0.03771                           | 0.18000                           | 0.04973                           | 0.24160                           | 0.01043                           | 0.26055                           | 0.09994                 |
| <b>rs2869550</b>  | 0.06303                           | 0.02152                           | 0.15877                           | 0.01963                           | 0.07555                           | 0.14941                           | 0.08703                           | 0.14425                           | 0.01501*                          | 0.15287                           | 0.08570                 |
| <b>rs7198325</b>  | 0.09113                           | 0.19367                           | 0.03953                           | 0.05931                           | 0.18518                           | 0.01100                           | 0.02889                           | 0.27872                           | 0.25715                           | 0.00333                           | 0.11479                 |
| <b>rs4470161</b>  | 0.07818                           | 0.18528                           | 0.13241                           | 0.08502                           | 0.17175                           | 0.01589                           | 0.01104*                          | 0.24469                           | 0.21048                           | 0.02841                           | 0.11411                 |
| <b>rs1019118</b>  | 0.14512                           | 0.18108                           | 0.18284                           | 0.15749                           | 0.10161                           | 0.00191                           | 0.01339                           | 0.10697                           | 0.11458                           | 0.00519*                          | 0.09998                 |
| <b>rs17246021</b> | 0.07581                           | 0.05279                           | 0.22181                           | 0.05652                           | 0.10004                           | 0.16398                           | 0.10776                           | 0.17468                           | 0.02013                           | 0.19030                           | 0.11638                 |
| <b>rs12936629</b> | 0.01998                           | 0.01411                           | 0.21765                           | 0.02500                           | 0.04430                           | 0.17194                           | 0.06143                           | 0.23753                           | 0.04229                           | 0.26252                           | 0.10967                 |
| <b>rs221308</b>   | 0.14085                           | 0.16834                           | 0.18475                           | 0.17292                           | 0.13002                           | 0.05482                           | 0.05504                           | 0.05435                           | 0.04844                           | 0.01656*                          | 0.09930                 |
| <b>rs6015771</b>  | 0.15009                           | 0.13830                           | 0.20477                           | 0.16103                           | 0.17017                           | 0.12877                           | 0.11655                           | 0.05072                           | 0.01474                           | 0.03421                           | 0.11694                 |
| <b>rs1013001</b>  | 0.07874                           | 0.06691                           | 0.21432                           | 0.06699                           | 0.10817                           | 0.14339                           | 0.09917                           | 0.15517                           | 0.03093                           | 0.16044                           | 0.11242                 |
| <b>rs2834567</b>  | 0.14717                           | 0.15771                           | 0.20497                           | 0.16125                           | 0.11549                           | 0.08042                           | 0.10691                           | 0.02358                           | 0.03044                           | 0.01768                           | 0.10456                 |
| <b>rs440431</b>   | 0.11581                           | 0.18229                           | 0.09782                           | 0.09714                           | 0.14255                           | 0.01190*                          | 0.01318*                          | 0.21730                           | 0.19426                           | 0.01947*                          | 0.10026                 |
| <b>rs1000472</b>  | 0.16861                           | 0.18993                           | 0.20436                           | 0.19454                           | 0.08235                           | 0.04620                           | 0.05654                           | 0.03668                           | 0.04488                           | 0.00699                           | 0.10311                 |

**Supplemental Table S2.1: F<sub>ST</sub> Statistic for the Setser80.**

All ten possible pairwise F<sub>ST</sub> statistics for the five populations from the GOAL study. Numbers listed in bold denote F<sub>ST</sub> values above the 0.15 F<sub>ST</sub> threshold. Abbreviations used: SNP = single nucleotide polymorphism, HUR = Honduras, DOM = Dominican Republic, COL = Colombia, CUB = Cuba, PUR = Puerto Rico. Asterisk\* denotes values less than 0.

**Supplemental Table S2.2: Country Attributable Mean  $F_{ST}$  for the Setser80**

| <b>Chr</b> | <b>SNP</b> | <b>Position<br/>NCBI36/hg18</b> | <b>COL<br/><math>F_{ST}</math></b> | <b>CUB<br/><math>F_{ST}</math></b> | <b>DOM<br/><math>F_{ST}</math></b> | <b>HUR<br/><math>F_{ST}</math></b> | <b>PUR<br/><math>F_{ST}</math></b> | <b>1<sup>st</sup><br/>Country</b> | <b>2<sup>nd</sup><br/>Country</b> |
|------------|------------|---------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|-----------------------------------|-----------------------------------|
| 1          | rs12130873 | 14629827                        | 0.07193                            | 0.06421                            | 0.08298                            | 0.27834                            | 0.08626                            | HUR                               | PUR                               |
| 1          | rs6694897  | 25990886                        | 0.16554                            | 0.11347                            | 0.08868                            | 0.07282                            | 0.06608                            | COL                               | CUB                               |
| 1          | rs1570099  | 35071891                        | 0.14964                            | 0.09674                            | 0.08445                            | 0.06265                            | 0.06167                            | COL                               | CUB                               |
| 1          | rs10493701 | 81919788                        | 0.06695                            | 0.09218                            | 0.07675                            | 0.21506                            | 0.07748                            | HUR                               | CUB                               |
| 1          | rs1040424  | 208201005                       | 0.05155                            | 0.09197                            | 0.05697                            | 0.15556                            | 0.05755                            | HUR                               | CUB                               |
| 2          | rs10495889 | 41547938                        | 0.07187                            | 0.09780                            | 0.08401                            | 0.16222                            | 0.07677                            | HUR                               | CUB                               |
| 2          | rs17018313 | 80045074                        | 0.04948                            | 0.06075                            | 0.06979                            | 0.23693                            | 0.07268                            | HUR                               | PUR                               |
| 2          | rs7568419  | 177078765                       | 0.06454                            | 0.07581                            | 0.08859                            | 0.22077                            | 0.09977                            | HUR                               | PUR                               |
| 2          | rs11693873 | 197850455                       | 0.07454                            | 0.09289                            | 0.06614                            | 0.15274                            | 0.06672                            | HUR                               | CUB                               |
| 2          | rs4433950  | 239329179                       | 0.05052                            | 0.06625                            | 0.07269                            | 0.23298                            | 0.07530                            | HUR                               | PUR                               |
| 3          | rs9853146  | 12274313                        | 0.14598                            | 0.14106                            | 0.08171                            | 0.09794                            | 0.07303                            | COL                               | CUB                               |
| 3          | rs4685443  | 17205888                        | 0.13890                            | 0.07009                            | 0.07537                            | 0.05645                            | 0.05042                            | COL                               | DOM                               |
| 3          | rs259425   | 22041732                        | 0.16228                            | 0.08050                            | 0.10253                            | 0.07050                            | 0.06818                            | COL                               | DOM                               |
| 3          | rs2356298  | 51509761                        | 0.09430                            | 0.11188                            | 0.08027                            | 0.16715                            | 0.07657                            | HUR                               | CUB                               |
| 3          | rs11719358 | 139267512                       | 0.12963                            | 0.09319                            | 0.12198                            | 0.09278                            | 0.07028                            | COL                               | DOM                               |
| 3          | rs9857908  | 139318436                       | 0.16058                            | 0.12885                            | 0.15183                            | 0.14021                            | 0.10139                            | COL                               | DOM                               |
| 3          | rs1586861  | 140541186                       | 0.16976                            | 0.14602                            | 0.07425                            | 0.08544                            | 0.07283                            | COL                               | CUB                               |
| 4          | rs3910480  | 161880988                       | 0.14157                            | 0.07363                            | 0.05788                            | 0.05506                            | 0.06986                            | COL                               | CUB                               |
| 5          | rs1366363  | 29985377                        | 0.07655                            | 0.11796                            | 0.08198                            | 0.20338                            | 0.08116                            | HUR                               | CUB                               |
| 5          | rs3733838  | 42757753                        | 0.06120                            | 0.08908                            | 0.07092                            | 0.21079                            | 0.07717                            | HUR                               | CUB                               |
| 5          | rs1438745  | 85249943                        | 0.06592                            | 0.06091                            | 0.22267                            | 0.07837                            | 0.09995                            | DOM                               | PUR                               |
| 5          | rs16902270 | 85859474                        | 0.16480                            | 0.10138                            | 0.08442                            | 0.06625                            | 0.06390                            | COL                               | CUB                               |
| 5          | rs871234   | 174418204                       | 0.09797                            | 0.06186                            | 0.20338                            | 0.09484                            | 0.08261                            | DOM                               | COL                               |
| 5          | rs692713   | 176186041                       | 0.11848                            | 0.06475                            | 0.20640                            | 0.09925                            | 0.07685                            | DOM                               | COL                               |
| 5          | rs4608884  | 178979749                       | 0.14133                            | 0.10817                            | 0.10083                            | 0.07143                            | 0.08111                            | COL                               | CUB                               |

| <b>Chr</b> | <b>SNP</b> | <b>Position<br/>NCBI36/hg18</b> | <b>COL<br/>F<sub>ST</sub></b> | <b>CUB<br/>F<sub>ST</sub></b> | <b>DOM<br/>F<sub>ST</sub></b> | <b>HUR<br/>F<sub>ST</sub></b> | <b>PUR<br/>F<sub>ST</sub></b> | <b>1<sup>st</sup><br/>Country</b> | <b>2<sup>nd</sup><br/>Country</b> |
|------------|------------|---------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-----------------------------------|-----------------------------------|
| 5          | rs190592   | 179263370                       | 0.13208                       | 0.06942                       | 0.08445                       | 0.09200                       | 0.07915                       | DOM                               | COL                               |
| 6          | rs6596807  | 1267033                         | 0.07007                       | 0.09021                       | 0.07752                       | 0.20170                       | 0.08519                       | HUR                               | CUB                               |
| 6          | rs9392285  | 1310265                         | 0.07697                       | 0.11633                       | 0.17955                       | 0.19772                       | 0.09253                       | HUR                               | CUB                               |
| 6          | rs9501948  | 3156786                         | 0.11352                       | 0.06213                       | 0.08503                       | 0.09332                       | 0.08925                       | DOM                               | COL                               |
| 6          | rs1329521  | 47840137                        | 0.16820                       | 0.13484                       | 0.07866                       | 0.07572                       | 0.07269                       | COL                               | CUB                               |
| 6          | rs17745021 | 73769618                        | 0.04929                       | 0.04948                       | 0.07148                       | 0.22153                       | 0.08811                       | HUR                               | PUR                               |
| 6          | rs3777908  | 111980345                       | 0.05423                       | 0.05424                       | 0.08166                       | 0.23860                       | 0.07442                       | HUR                               | PUR                               |
| 6          | rs17087570 | 156949439                       | 0.06907                       | 0.11388                       | 0.10892                       | 0.25097                       | 0.08757                       | HUR                               | CUB                               |
| 6          | rs4709836  | 164636380                       | 0.14426                       | 0.08011                       | 0.19199                       | 0.06399                       | 0.06473                       | COL                               | DOM                               |
| 7          | rs12536738 | 95370320                        | 0.08904                       | 0.11446                       | 0.07656                       | 0.18567                       | 0.07744                       | HUR                               | CUB                               |
| 7          | rs10953750 | 113424187                       | 0.06635                       | 0.09920                       | 0.09400                       | 0.15649                       | 0.06471                       | HUR                               | CUB                               |
| 7          | rs2352479  | 137589866                       | 0.12823                       | 0.07748                       | 0.09892                       | 0.06024                       | 0.07394                       | COL                               | DOM                               |
| 7          | rs17480133 | 145414794                       | 0.19653                       | 0.11315                       | 0.08312                       | 0.07909                       | 0.07932                       | COL                               | CUB                               |
| 8          | rs766382   | 60170978                        | 0.11356                       | 0.11474                       | 0.08048                       | 0.08270                       | 0.06200                       | CUB                               | COL                               |
| 8          | rs1588459  | 142143028                       | 0.17997                       | 0.10768                       | 0.08445                       | 0.07261                       | 0.06607                       | COL                               | CUB                               |
| 9          | rs6474712  | 12405123                        | 0.08651                       | 0.06513                       | 0.23846                       | 0.07985                       | 0.07285                       | DOM                               | COL                               |
| 9          | rs880397   | 12451349                        | 0.08143                       | 0.07307                       | 0.22029                       | 0.07348                       | 0.06905                       | DOM                               | COL                               |
| 9          | rs10981894 | 115508956                       | 0.10335                       | 0.13274                       | 0.11006                       | 0.10909                       | 0.12488                       | CUB                               | PUR                               |
| 10         | rs2008617  | 34499344                        | 0.09387                       | 0.07103                       | 0.28372                       | 0.09504                       | 0.10056                       | DOM                               | PUR                               |
| 10         | rs16912280 | 59880729                        | 0.10208                       | 0.07394                       | 0.26838                       | 0.10046                       | 0.08559                       | DOM                               | COL                               |
| 10         | rs1259603  | 76814933                        | 0.07241                       | 0.08035                       | 0.09140                       | 0.23429                       | 0.10085                       | HUR                               | PUR                               |
| 10         | rs16932385 | 77017391                        | 0.07676                       | 0.09621                       | 0.10156                       | 0.27760                       | 0.10298                       | HUR                               | PUR                               |
| 10         | rs11189628 | 100230671                       | 0.15679                       | 0.10420                       | 0.06875                       | 0.06358                       | 0.06298                       | COL                               | CUB                               |
| 10         | rs17112705 | 101927184                       | 0.17958                       | 0.11240                       | 0.11049                       | 0.08208                       | 0.07515                       | COL                               | CUB                               |
| 11         | rs1849352  | 61578905                        | 0.07144                       | 0.08943                       | 0.08481                       | 0.18204                       | 0.07139                       | HUR                               | CUB                               |
| 11         | rs2878712  | 132343165                       | 0.06861                       | 0.07834                       | 0.08526                       | 0.22463                       | 0.09477                       | HUR                               | PUR                               |
| 12         | rs10840730 | 17553352                        | 0.05327                       | 0.07914                       | 0.07647                       | 0.22546                       | 0.07359                       | HUR                               | CUB                               |
| 12         | rs2051827  | 46242298                        | 0.15890                       | 0.11428                       | 0.07672                       | 0.10426                       | 0.10131                       | COL                               | CUB                               |
| 12         | rs12146822 | 66857933                        | 0.12117                       | 0.15467                       | 0.08795                       | 0.18173                       | 0.09100                       | HUR                               | CUB                               |
| 12         | rs7310083  | 68347053                        | 0.13093                       | 0.06771                       | 0.07504                       | 0.08601                       | 0.21863                       | PUR                               | COL                               |

| <b>Chr</b> | <b>SNP</b> | <b>Position<br/>NCBI36/hg18</b> | <b>COL<br/>F<sub>ST</sub></b> | <b>CUB<br/>F<sub>ST</sub></b> | <b>DOM<br/>F<sub>ST</sub></b> | <b>HUR<br/>F<sub>ST</sub></b> | <b>PUR<br/>F<sub>ST</sub></b> | <b>1<sup>st</sup><br/>Country</b> | <b>2<sup>nd</sup><br/>Country</b> |
|------------|------------|---------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-----------------------------------|-----------------------------------|
| 12         | rs1967232  | 115531283                       | 0.14664                       | 0.10884                       | 0.06467                       | 0.05950                       | 0.05823                       | COL                               | CUB                               |
| 12         | rs6486527  | 129231989                       | 0.11961                       | 0.08767                       | 0.10008                       | 0.06631                       | 0.09289                       | COL                               | DOM                               |
| 13         | rs9569702  | 56996265                        | 0.05866                       | 0.08848                       | 0.05589                       | 0.15029                       | 0.06130                       | HUR                               | CUB                               |
| 13         | rs4341647  | 77872864                        | 0.06194                       | 0.09511                       | 0.06630                       | 0.17141                       | 0.07666                       | HUR                               | CUB                               |
| 13         | rs9556940  | 97818960                        | 0.16213                       | 0.10425                       | 0.06142                       | 0.06003                       | 0.05929                       | COL                               | CUB                               |
| 14         | rs1957572  | 67806224                        | 0.07067                       | 0.09433                       | 0.09263                       | 0.17737                       | 0.07417                       | HUR                               | CUB                               |
| 14         | rs178384   | 79252594                        | 0.10392                       | 0.12113                       | 0.07711                       | 0.09150                       | 0.07028                       | CUB                               | COL                               |
| 14         | rs12434466 | 96394042                        | 0.06368                       | 0.10147                       | 0.07296                       | 0.20049                       | 0.07356                       | HUR                               | CUB                               |
| 14         | rs17094860 | 96560669                        | 0.07280                       | 0.10820                       | 0.09627                       | 0.23314                       | 0.08505                       | HUR                               | CUB                               |
| 14         | rs12435621 | 97182294                        | 0.07145                       | 0.08070                       | 0.09498                       | 0.31859                       | 0.09892                       | HUR                               | PUR                               |
| 14         | rs12431505 | 97188301                        | 0.09419                       | 0.11915                       | 0.12273                       | 0.39018                       | 0.12929                       | HUR                               | PUR                               |
| 14         | rs1462266  | 97319155                        | 0.06255                       | 0.09924                       | 0.08161                       | 0.24835                       | 0.08378                       | HUR                               | CUB                               |
| 14         | rs17097005 | 98013715                        | 0.05485                       | 0.07184                       | 0.07159                       | 0.22045                       | 0.08099                       | HUR                               | PUR                               |
| 15         | rs2869550  | 76768056                        | 0.06574                       | 0.09375                       | 0.05658                       | 0.15132                       | 0.06113                       | HUR                               | CUB                               |
| 16         | rs7198325  | 12572650                        | 0.09591                       | 0.07905                       | 0.22868                       | 0.08314                       | 0.08717                       | DOM                               | COL                               |
| 16         | rs4470161  | 78627029                        | 0.12023                       | 0.06370                       | 0.20305                       | 0.10535                       | 0.07822                       | DOM                               | COL                               |
| 17         | rs1019118  | 52103046                        | 0.16663                       | 0.06551                       | 0.12606                       | 0.07163                       | 0.07007                       | COL                               | DOM                               |
| 17         | rs17246021 | 67616034                        | 0.10173                       | 0.11189                       | 0.08691                       | 0.18769                       | 0.09368                       | HUR                               | CUB                               |
| 17         | rs12936629 | 67854083                        | 0.06918                       | 0.07441                       | 0.08456                       | 0.22241                       | 0.09781                       | HUR                               | PUR                               |
| 20         | rs221308   | 34709812                        | 0.16671                       | 0.09518                       | 0.10029                       | 0.06934                       | 0.06496                       | COL                               | DOM                               |
| 20         | rs6015771  | 58532301                        | 0.16355                       | 0.14140                       | 0.09348                       | 0.10462                       | 0.08163                       | COL                               | CUB                               |
| 21         | rs1013001  | 14581125                        | 0.10674                       | 0.10737                       | 0.09030                       | 0.16833                       | 0.08938                       | HUR                               | CUB                               |
| 21         | rs2834567  | 34950995                        | 0.16778                       | 0.11250                       | 0.08180                       | 0.08166                       | 0.07907                       | COL                               | CUB                               |
| 21         | rs440431   | 42905781                        | 0.12326                       | 0.05832                       | 0.18410                       | 0.07094                       | 0.06469                       | DOM                               | COL                               |
| 21         | rs1000472  | 43602306                        | 0.18936                       | 0.08843                       | 0.08846                       | 0.07356                       | 0.07574                       | COL                               | DOM                               |

**Supplemental Table S2.2: Country Attributable Mean  $F_{ST}$  for the Setser80.** Each country attributable mean  $F_{ST}$  is calculated by averaging the four population comparisons that have one country in common (e.g. HUR vs. DOM, HUR vs. COL, HUR vs. CUB, and HUR vs. PUR). The highest country  $F_{ST}$  is the 1<sup>st</sup> country attributable mean  $F_{ST}$  value of the five populations and is shown in bold. The next highest country  $F_{ST}$  is the 2<sup>nd</sup> country attributable mean  $F_{ST}$  and is shown in italics. Abbreviations used: Chr = chromosome, SNP = single nucleotide polymorphism, HUR = Honduras, DOM = Dominican Republic, COL = Colombia, CUB = Cuba, and PUR = Puerto Rico.

**Supplemental Table S2.3: Summary of the  $F_{ST}$  Values of the Setser Panel SNPs**

|                                                      | <b>Setser234</b> | <b>Setser80</b> |
|------------------------------------------------------|------------------|-----------------|
| <b><math>F_{ST} \geq 0.15</math></b>                 | 769              | 266             |
| <b><math>F_{ST} &lt; 0.15</math></b>                 | 1571             | 534             |
| <b>Total</b>                                         | 2340             | 800             |
| <b>Proportion <math>F_{ST} \geq 0.15</math></b>      | 32.9%            | 33.3%           |
| <b>HUR vs. DOM <math>F_{ST}</math></b>               | 0.19507          | 0.16059         |
| <b>HUR vs. COL <math>F_{ST}</math></b>               | 0.13102          | 0.17113         |
| <b>HUR vs. PUR <math>F_{ST}</math></b>               | 0.13091          | 0.12042         |
| <b>HUR vs. CUB <math>F_{ST}</math></b>               | 0.09967          | 0.11534         |
| <b>DOM vs. CUB <math>F_{ST}</math></b>               | 0.09459          | 0.10858         |
| <b>DOM vs. PUR <math>F_{ST}</math></b>               | 0.08400          | 0.05933         |
| <b>DOM vs. COL <math>F_{ST}</math></b>               | 0.07687          | 0.09258         |
| <b>COL vs. CUB <math>F_{ST}</math></b>               | 0.05105          | 0.08509         |
| <b>COL vs. PUR <math>F_{ST}</math></b>               | 0.04137          | 0.07991         |
| <b>CUB vs. PUR <math>F_{ST}</math></b>               | 0.03994          | 0.06415         |
| <b>Overall 1<sup>st</sup></b>                        |                  |                 |
| <b>Country Attributable Mean <math>F_{ST}</math></b> | 0.19124          | 0.19194         |
| <b>HUR <math>F_{ST}</math></b>                       | 0.13974          | 0.14187         |
| <b>DOM <math>F_{ST}</math></b>                       | 0.11240          | 0.10527         |
| <b>COL <math>F_{ST}</math></b>                       | 0.07477          | 0.10718         |
| <b>PUR <math>F_{ST}</math></b>                       | 0.07389          | 0.08095         |
| <b>CUB <math>F_{ST}</math></b>                       | 0.07116          | 0.09329         |

**Supplemental Table S2.3: Summary of  $F_{ST}$  Values of the Setser Panel SNPs**

Description of the two Setser panels from the top down by the amount of  $F_{ST} \geq 0.15$ .

First = number of eligible SNPs after the data from GOAL study was filtered for quality.

Second = mean  $F_{ST}$  value for each pairwise comparison on the selected SNPs.

Third = mean  $F_{ST}$  values for the 4 pairwise comparisons that are combined to create the 1<sup>st</sup> country attributable mean  $F_{ST}$  for each population. Abbreviations used: HUR = Honduras, DOM = Dominican Republic, COL = Colombia, CUB = Cuba, and PUR = Puerto Rico.

**Supplemental Table S2.4: Allele Frequencies for the Setser80**

|                     | <b>Loci<br/>Missing<br/>Maj./Min.</b> | <b>rs12130873</b><br>4<br>G/T | <b>rs6694897</b><br>0<br>A/G | <b>rs1570099</b><br>0<br>G/C  | <b>rs10493701</b><br>1<br>G/A | <b>rs1040424</b><br>1<br>A/G | <b>rs10495889</b><br>2<br>A/C |
|---------------------|---------------------------------------|-------------------------------|------------------------------|-------------------------------|-------------------------------|------------------------------|-------------------------------|
| <b>ALL</b><br>n=160 | Major Freq<br>Minor Freq              | 0.98125<br>0.01875            | 0.40549<br>0.59451           | 0.69207<br>0.30793            | 0.84663<br>0.15337            | 0.87117<br>0.12883           | 0.77778<br>0.22222            |
| <b>HUR</b><br>n=13  | Major Freq<br>Minor Freq              | 0.80769<br>0.19231            | 0.5<br>0.5                   | 0.65385<br>0.34615            | 0.46154<br>0.53846            | 0.57692<br>0.42308           | 0.42308<br>0.57692            |
| <b>DOM</b><br>n=21  | Major Freq<br>Minor Freq              | 1<br>0                        | 0.23810<br>0.76190           | 0.85714<br>0.14286            | 0.90476<br>0.09524            | 0.90000<br>0.10000           | 0.90476<br>0.09524            |
| <b>COL</b><br>n=53  | Major Freq<br>Minor Freq              | 1<br>0                        | 0.63208<br>0.36792           | 0.49057<br>0.50943            | 0.79245<br>0.20755            | 0.83019<br>0.16981           | 0.69811<br>0.30189            |
| <b>CUB</b><br>n=55  | Major Freq<br>Minor Freq              | 0.99091<br>0.00909            | 0.24545<br>0.75455           | 0.82727<br>0.17273            | 0.94444<br>0.05556            | 0.97273<br>0.02727           | 0.90566<br>0.09434            |
| <b>PUR</b><br>n=18  | Major Freq<br>Minor Freq              | 1<br>0                        | 0.41667<br>0.58333           | 0.63889<br>0.36111            | 0.88889<br>0.11111            | 0.86111<br>0.13889           | 0.69444<br>0.30556            |
|                     | <b>Loci<br/>Missing<br/>Maj./Min.</b> | <b>rs17018313</b><br>0<br>A/G | <b>rs7568419</b><br>0<br>A/C | <b>rs11693873</b><br>1<br>C/T | <b>rs4433950</b><br>0<br>C/T  | <b>rs9853146</b><br>1<br>A/G | <b>rs4685443</b><br>0<br>T/C  |
| <b>ALL</b><br>n=160 | Major Freq<br>Minor Freq              | 0.97866<br>0.02134            | 0.71951<br>0.28049           | 0.39877<br>0.60123            | 0.96951<br>0.03049            | 0.43865<br>0.56135           | 0.81707<br>0.18293            |
| <b>HUR</b><br>n=13  | Major Freq<br>Minor Freq              | 0.80769<br>0.19231            | 0.23077<br>0.76923           | 0.80769<br>0.19231            | 0.76923<br>0.23077            | 0.65385<br>0.34615           | 0.88462<br>0.11538            |
| <b>DOM</b><br>n=21  | Major Freq<br>Minor Freq              | 1<br>0                        | 0.83333<br>0.16667           | 0.325<br>0.675                | 1<br>0                        | 0.325<br>0.675               | 0.95238<br>0.04762            |
| <b>COL</b><br>n=53  | Major Freq<br>Minor Freq              | 0.98113<br>0.01887            | 0.76415<br>0.23585           | 0.50943<br>0.49057            | 0.96226<br>0.03774            | 0.64151<br>0.35849           | 0.65094<br>0.34906            |
| <b>CUB</b><br>n=55  | Major Freq<br>Minor Freq              | 1<br>0                        | 0.8<br>0.2                   | 0.25455<br>0.74545            | 1<br>0                        | 0.24545<br>0.75455           | 0.9<br>0.1                    |
| <b>PUR</b><br>n=18  | Major Freq<br>Minor Freq              | 1<br>0                        | 0.55556<br>0.44444           | 0.33333<br>0.66667            | 1<br>0                        | 0.41667<br>0.58333           | 0.80556<br>0.19444            |
|                     | <b>Loci<br/>Missing<br/>Maj./Min.</b> | <b>rs259425</b><br>0<br>A/G   | <b>rs2356298</b><br>0<br>T/C | <b>rs11719358</b><br>1<br>G/A | <b>rs9857908</b><br>0<br>C/T  | <b>rs1586861</b><br>2<br>T/C | <b>rs3910480</b><br>7<br>G/A  |
| <b>ALL</b><br>n=160 | Major Freq<br>Minor Freq              | 0.70122<br>0.29878            | 0.88720<br>0.11280           | 0.61043<br>0.38957            | 0.56098<br>0.43902            | 0.70988<br>0.29012           | 0.71656<br>0.28344            |
| <b>HUR</b><br>n=13  | Major Freq<br>Minor Freq              | 0.61538<br>0.38462            | 0.61538<br>0.38462           | 0.80769<br>0.19231            | 0.84615<br>0.15385            | 0.57692<br>0.42308           | 0.76923<br>0.23077            |

|                     |                                                   |                              |                              |                               |                               |                              |                              |
|---------------------|---------------------------------------------------|------------------------------|------------------------------|-------------------------------|-------------------------------|------------------------------|------------------------------|
| <b>DOM</b><br>n=21  | Major Freq                                        | 0.90476                      | 0.95238                      | 0.35714                       | 0.28571                       | 0.76190                      | 0.8                          |
|                     | Minor Freq                                        | 0.09524                      | 0.04762                      | 0.64286                       | 0.71429                       | 0.23810                      | 0.2                          |
|                     | <b>Loci</b><br><b>Missing</b><br><b>Maj./Min.</b> | <b>rs259425</b><br>0<br>A/G  | <b>rs2356298</b><br>0<br>T/C | <b>rs11719358</b><br>1<br>G/A | <b>rs9857908</b><br>0<br>C/T  | <b>rs1586861</b><br>2<br>T/C | <b>rs3910480</b><br>7<br>G/A |
| <b>COL</b><br>n=53  | Major Freq                                        | 0.49057                      | 0.80189                      | 0.79808                       | 0.76415                       | 0.5                          | 0.51961                      |
|                     | Minor Freq                                        | 0.50943                      | 0.19811                      | 0.20192                       | 0.23585                       | 0.5                          | 0.48039                      |
| <b>CUB</b><br>n=55  | Major Freq                                        | 0.8                          | 0.99091                      | 0.48182                       | 0.4                           | 0.89091                      | 0.81481                      |
|                     | Minor Freq                                        | 0.2                          | 0.00909                      | 0.51818                       | 0.6                           | 0.10909                      | 0.18519                      |
| <b>PUR</b><br>n=18  | Major Freq                                        | 0.77778                      | 0.91667                      | 0.63889                       | 0.63889                       | 0.72222                      | 0.86667                      |
|                     | Minor Freq                                        | 0.22222                      | 0.08333                      | 0.36111                       | 0.36111                       | 0.27778                      | 0.13333                      |
|                     | <b>Loci</b><br><b>Missing</b><br><b>Maj./Min.</b> | <b>rs1366363</b><br>2<br>G/A | <b>rs3733838</b><br>1<br>A/G | <b>rs1438745</b><br>0<br>T/A  | <b>rs16902270</b><br>5<br>C/A | <b>rs871234</b><br>0<br>T/G  | <b>rs692713</b><br>0<br>A/G  |
| <b>ALL</b><br>n=160 | Major Freq                                        | 0.81173                      | 0.93558                      | 0.85976                       | 0.86164                       | 0.88415                      | 0.71341                      |
|                     | Minor Freq                                        | 0.18827                      | 0.06442                      | 0.14024                       | 0.13836                       | 0.11585                      | 0.28659                      |
| <b>HUR</b><br>n=13  | Major Freq                                        | 0.42308                      | 0.66667                      | 0.88462                       | 0.83333                       | 1                            | 0.88462                      |
|                     | Minor Freq                                        | 0.57692                      | 0.33333                      | 0.11538                       | 0.16667                       | 0                            | 0.11538                      |
| <b>DOM</b><br>n=21  | Major Freq                                        | 0.875                        | 0.95238                      | 0.54762                       | 0.975                         | 0.61905                      | 0.33333                      |
|                     | Minor Freq                                        | 0.125                        | 0.04762                      | 0.45238                       | 0.025                         | 0.38095                      | 0.66667                      |
| <b>COL</b><br>n=53  | Major Freq                                        | 0.74528                      | 0.90566                      | 0.90566                       | 0.70192                       | 0.97170                      | 0.85849                      |
|                     | Minor Freq                                        | 0.25472                      | 0.09434                      | 0.09434                       | 0.29808                       | 0.02830                      | 0.14151                      |
| <b>CUB</b><br>n=55  | Major Freq                                        | 0.94444                      | 1                            | 0.89091                       | 0.96226                       | 0.9                          | 0.72727                      |
|                     | Minor Freq                                        | 0.05556                      | 0                            | 0.10909                       | 0.03774                       | 0.1                          | 0.27273                      |
| <b>PUR</b><br>n=18  | Major Freq                                        | 0.77778                      | 0.97222                      | 1                             | 0.88889                       | 0.80556                      | 0.66667                      |
|                     | Minor Freq                                        | 0.22222                      | 0.02778                      | 0                             | 0.11111                       | 0.19444                      | 0.33333                      |
|                     | <b>Loci</b><br><b>Missing</b><br><b>Maj./Min.</b> | <b>rs4608884</b><br>0<br>G/A | <b>rs190592</b><br>0<br>A/G  | <b>rs6596807</b><br>0<br>A/G  | <b>rs9392285</b><br>0<br>C/G  | <b>rs9501948</b><br>0<br>T/C | <b>rs1329521</b><br>0<br>G/A |
| <b>ALL</b><br>n=160 | Major Freq                                        | 0.85366                      | 0.87500                      | 0.90854                       | 0.88720                       | 0.83232                      | 0.49695                      |
|                     | Minor Freq                                        | 0.14634                      | 0.12500                      | 0.09146                       | 0.11280                       | 0.16768                      | 0.50305                      |
| <b>HUR</b><br>n=13  | Major Freq                                        | 0.88462                      | 0.96154                      | 0.61538                       | 0.57692                       | 0.96154                      | 0.57692                      |
|                     | Minor Freq                                        | 0.11538                      | 0.03846                      | 0.38462                       | 0.42308                       | 0.03846                      | 0.42308                      |
| <b>DOM</b><br>n=21  | Major Freq                                        | 1                            | 0.59524                      | 0.97619                       | 0.88095                       | 0.54762                      | 0.35714                      |
|                     | Minor Freq                                        | 0                            | 0.40476                      | 0.02381                       | 0.11905                       | 0.45238                      | 0.64286                      |
| <b>COL</b><br>n=53  | Major Freq                                        | 0.70755                      | 0.99057                      | 0.85849                       | 0.83019                       | 0.95283                      | 0.72642                      |
|                     | Minor Freq                                        | 0.29245                      | 0.00943                      | 0.14151                       | 0.16981                       | 0.04717                      | 0.27358                      |
| <b>CUB</b><br>n=55  | Major Freq                                        | 0.96364                      | 0.89091                      | 0.98182                       | 0.99091                       | 0.79091                      | 0.30909                      |
|                     | Minor Freq                                        | 0.03636                      | 0.10909                      | 0.01818                       | 0.00909                       | 0.20909                      | 0.69091                      |
| <b>PUR</b><br>n=18  | Major Freq                                        | 0.75                         | 0.83333                      | 0.97222                       | 0.97222                       | 0.94444                      | 0.55556                      |
|                     | Minor Freq                                        | 0.25                         | 0.16667                      | 0.02778                       | 0.02778                       | 0.05556                      | 0.44444                      |

|                     | <b>Loci</b><br><b>Missing</b><br><b>Maj./Min.</b> | <b>rs17745021</b><br>0<br>T/C | <b>rs3777908</b><br>0<br>C/T  | <b>rs17087570</b><br>0<br>C/T | <b>rs4709836</b><br>0<br>C/T | <b>rs12536738</b><br>7<br>C/T | <b>rs10953750</b><br>0<br>A/G |
|---------------------|---------------------------------------------------|-------------------------------|-------------------------------|-------------------------------|------------------------------|-------------------------------|-------------------------------|
| <b>ALL</b><br>n=160 | Major Freq                                        | 0.90549                       | 0.97866                       | 0.91159                       | 0.73476                      | 0.85987                       | 0.82622                       |
|                     | Minor Freq                                        | 0.09451                       | 0.02134                       | 0.08841                       | 0.26524                      | 0.14013                       | 0.17378                       |
| <b>HUR</b><br>n=13  | Major Freq                                        | 0.57692                       | 0.80769                       | 0.57692                       | 0.73077                      | 0.53846                       | 0.5                           |
|                     | Minor Freq                                        | 0.42308                       | 0.19231                       | 0.42308                       | 0.26923                      | 0.46154                       | 0.5                           |
| <b>DOM</b><br>n=21  | Major Freq                                        | 0.97619                       | 1                             | 0.90476                       | 0.95238                      | 0.90476                       | 0.85714                       |
|                     | Minor Freq                                        | 0.02381                       | 0                             | 0.09524                       | 0.04762                      | 0.09524                       | 0.14286                       |
| <b>COL</b><br>n=53  | Major Freq                                        | 0.91509                       | 0.99057                       | 0.88679                       | 0.54717                      | 0.77174                       | 0.75472                       |
|                     | Minor Freq                                        | 0.08491                       | 0.00943                       | 0.11321                       | 0.45283                      | 0.22826                       | 0.24528                       |
| <b>CUB</b><br>n=55  | Major Freq                                        | 0.91818                       | 0.99091                       | 1                             | 0.83636                      | 0.97273                       | 0.94545                       |
|                     | Minor Freq                                        | 0.08182                       | 0.00909                       | 0                             | 0.16364                      | 0.02727                       | 0.05455                       |
| <b>PUR</b><br>n=18  | Major Freq                                        | 1                             | 1                             | 0.94444                       | 0.66667                      | 0.88889                       | 0.86111                       |
|                     | Minor Freq                                        | 0                             | 0                             | 0.05556                       | 0.33333                      | 0.11111                       | 0.13889                       |
|                     | <b>Loci</b><br><b>Missing</b><br><b>Maj./Min.</b> | <b>rs2352479</b><br>1<br>C/T  | <b>rs17480133</b><br>0<br>C/G | <b>rs766382</b><br>0<br>T/G   | <b>rs1588459</b><br>0<br>G/A | <b>rs6474712</b><br>0<br>A/G  | <b>rs880397</b><br>0<br>G/C   |
| <b>ALL</b><br>n=160 | Major Freq                                        | 0.45706                       | 0.40854                       | 0.80183                       | 0.85976                      | 0.92073                       | 0.96037                       |
|                     | Minor Freq                                        | 0.54294                       | 0.59146                       | 0.19817                       | 0.14024                      | 0.07927                       | 0.03963                       |
| <b>HUR</b><br>n=13  | Major Freq                                        | 0.46154                       | 0.46154                       | 0.96154                       | 0.92308                      | 0.96154                       | 0.96154                       |
|                     | Minor Freq                                        | 0.53846                       | 0.53846                       | 0.03846                       | 0.07692                      | 0.03846                       | 0.03846                       |
| <b>DOM</b><br>n=21  | Major Freq                                        | 0.23810                       | 0.23810                       | 0.66667                       | 0.95238                      | 0.66667                       | 0.78571                       |
|                     | Minor Freq                                        | 0.76190                       | 0.76190                       | 0.33333                       | 0.04762                      | 0.33333                       | 0.21429                       |
| <b>COL</b><br>n=53  | Major Freq                                        | 0.65385                       | 0.66038                       | 0.94340                       | 0.68868                      | 0.98113                       | 1                             |
|                     | Minor Freq                                        | 0.34615                       | 0.33962                       | 0.05660                       | 0.31132                      | 0.01887                       | 0                             |
| <b>CUB</b><br>n=55  | Major Freq                                        | 0.33636                       | 0.26364                       | 0.66364                       | 0.96364                      | 0.94545                       | 0.99091                       |
|                     | Minor Freq                                        | 0.66364                       | 0.73636                       | 0.33636                       | 0.03636                      | 0.05455                       | 0.00909                       |
| <b>PUR</b><br>n=18  | Major Freq                                        | 0.61111                       | 0.33333                       | 0.83333                       | 0.86111                      | 0.94444                       | 0.97222                       |
|                     | Minor Freq                                        | 0.38889                       | 0.66667                       | 0.16667                       | 0.13889                      | 0.05556                       | 0.02778                       |
|                     | <b>Loci</b><br><b>Missing</b><br><b>Maj./Min.</b> | <b>rs10981894</b><br>8<br>G/A | <b>rs2008617</b><br>4<br>A/T  | <b>rs16912280</b><br>0<br>G/A | <b>rs1259603</b><br>0<br>C/A | <b>rs16932385</b><br>0<br>T/G | <b>rs11189628</b><br>0<br>T/C |
| <b>ALL</b><br>n=160 | Major Freq                                        | 0.75962                       | 0.92500                       | 0.92683                       | 0.90549                      | 0.90549                       | 0.68293                       |
|                     | Minor Freq                                        | 0.24038                       | 0.07500                       | 0.07317                       | 0.09451                      | 0.09451                       | 0.31707                       |
| <b>HUR</b><br>n=13  | Major Freq                                        | 0.57692                       | 0.96154                       | 1                             | 0.57692                      | 0.53846                       | 0.61538                       |
|                     | Minor Freq                                        | 0.42308                       | 0.03846                       | 0                             | 0.42308                      | 0.46154                       | 0.38462                       |
| <b>DOM</b><br>n=21  | Major Freq                                        | 0.92857                       | 0.65                          | 0.66667                       | 0.97619                      | 0.97619                       | 0.78571                       |
|                     | Minor Freq                                        | 0.07143                       | 0.35                          | 0.33333                       | 0.02381                      | 0.02381                       | 0.21429                       |

|            | <b>Loci</b>      | <b>rs10981894</b> | <b>rs2008617</b> | <b>rs16912280</b> | <b>rs1259603</b>  | <b>rs16932385</b> | <b>rs11189628</b> |
|------------|------------------|-------------------|------------------|-------------------|-------------------|-------------------|-------------------|
|            | <b>Missing</b>   | 8                 | 4                | 0                 | 0                 | 0                 | 0                 |
|            | <b>Maj./Min.</b> | G/A               | A/T              | G/A               | C/A               | T/G               | T/C               |
| <b>COL</b> | Major Freq       | 0.64              | 0.98113          | 0.99057           | 0.85849           | 0.86792           | 0.47170           |
| n=53       | Minor Freq       | 0.36              | 0.01887          | 0.00943           | 0.14151           | 0.13208           | 0.52830           |
| <b>CUB</b> | Major Freq       | 0.91509           | 0.93396          | 0.94545           | 0.96364           | 0.97273           | 0.82727           |
| n=55       | Minor Freq       | 0.08491           | 0.06604          | 0.05455           | 0.03636           | 0.02727           | 0.17273           |
| <b>PUR</b> | Major Freq       | 0.53333           | 1                | 0.94444           | 1                 | 0.97222           | 0.75              |
| n=18       | Minor Freq       | 0.46667           | 0                | 0.05556           | 0                 | 0.02778           | 0.25              |
|            | <b>Loci</b>      | <b>rs17112705</b> | <b>rs1849352</b> | <b>rs2878712</b>  | <b>rs10840730</b> | <b>rs2051827</b>  | <b>rs12146822</b> |
|            | <b>Missing</b>   | 0                 | 3                | 1                 | 1                 | 8                 | 0                 |
|            | <b>Maj./Min.</b> | T/C               | G/C              | G/A               | G/A               | G/A               | G/A               |
| ALL        | Major Freq       | 0.62805           | 0.72671          | 0.91718           | 0.92025           | 0.82692           | 0.82622           |
| n=160      | Minor Freq       | 0.37195           | 0.27329          | 0.08282           | 0.07975           | 0.17308           | 0.17378           |
| HUR        | Major Freq       | 0.53846           | 0.30769          | 0.61538           | 0.61538           | 0.65385           | 0.5               |
| n=13       | Minor Freq       | 0.46154           | 0.69231          | 0.38462           | 0.38462           | 0.34615           | 0.5               |
| DOM        | Major Freq       | 0.83333           | 0.85714          | 0.97619           | 0.97619           | 0.86842           | 0.85714           |
| n=21       | Minor Freq       | 0.16667           | 0.14286          | 0.02381           | 0.02381           | 0.13158           | 0.14286           |
| COL        | Major Freq       | 0.39623           | 0.64             | 0.875             | 0.91346           | 0.65625           | 0.70755           |
| n=53       | Minor Freq       | 0.60377           | 0.36             | 0.125             | 0.08654           | 0.34375           | 0.29245           |
| CUB        | Major Freq       | 0.77273           | 0.84545          | 0.97273           | 0.98182           | 0.94444           | 0.98182           |
| n=55       | Minor Freq       | 0.22727           | 0.15455          | 0.02727           | 0.01818           | 0.05556           | 0.01818           |
| PUR        | Major Freq       | 0.63889           | 0.72222          | 1                 | 0.88889           | 0.97222           | 0.86111           |
| n=18       | Minor Freq       | 0.36111           | 0.27778          | 0                 | 0.11111           | 0.02778           | 0.13889           |
|            | <b>Loci</b>      | <b>rs7310083</b>  | <b>rs1967232</b> | <b>rs6486527</b>  | <b>rs9569702</b>  | <b>rs4341647</b>  | <b>rs9556940</b>  |
|            | <b>Missing</b>   | 0                 | 1                | 13                | 0                 | 0                 | 0                 |
|            | <b>Maj./Min.</b> | G/T               | C/T              | T/C               | T/C               | C/T               | C/G               |
| ALL        | Major Freq       | 0.88110           | 0.81902          | 0.85762           | 0.77744           | 0.71951           | 0.85366           |
| n=160      | Minor Freq       | 0.11890           | 0.18098          | 0.14238           | 0.22256           | 0.28049           | 0.14634           |
| HUR        | Major Freq       | 0.92308           | 0.76923          | 0.88462           | 0.42308           | 0.30769           | 0.88462           |
| n=13       | Minor Freq       | 0.07692           | 0.23077          | 0.11538           | 0.57692           | 0.69231           | 0.11538           |
| DOM        | Major Freq       | 0.88095           | 0.9              | 0.68750           | 0.78571           | 0.66667           | 0.90476           |
| n=21       | Minor Freq       | 0.11905           | 0.1              | 0.31250           | 0.21429           | 0.33333           | 0.09524           |
| COL        | Major Freq       | 1                 | 0.65094          | 0.98039           | 0.70755           | 0.65094           | 0.68868           |

|       |                                       |                              |                               |                               |                               |                               |                               |
|-------|---------------------------------------|------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| n=53  | Minor Freq                            | 0                            | 0.34906                       | 0.01961                       | 0.29245                       | 0.34906                       | 0.31132                       |
| CUB   | Major Freq                            | 0.85455                      | 0.94545                       | 0.76                          | 0.9                           | 0.85455                       | 0.96364                       |
| n=55  | Minor Freq                            | 0.14545                      | 0.05455                       | 0.24                          | 0.1                           | 0.14545                       | 0.03636                       |
| PUR   | Major Freq                            | 0.58333                      | 0.86111                       | 1                             | 0.83333                       | 0.83333                       | 0.88889                       |
| n=18  | Minor Freq                            | 0.41667                      | 0.13889                       | 0                             | 0.16667                       | 0.16667                       | 0.11111                       |
|       | <b>Loci<br/>Missing<br/>Maj./Min.</b> | <b>rs1957572</b><br>2<br>T/C | <b>rs178384</b><br>0<br>C/T   | <b>rs12434466</b><br>0<br>A/G | <b>rs17094860</b><br>7<br>T/A | <b>rs12435621</b><br>1<br>T/G | <b>rs12431505</b><br>0<br>G/A |
| ALL   | Major Freq                            | 0.85802                      | 0.42683                       | 0.85671                       | 0.88535                       | 0.79448                       | 0.87195                       |
| n=160 | Minor Freq                            | 0.14198                      | 0.57317                       | 0.14329                       | 0.11465                       | 0.20552                       | 0.12805                       |
| HUR   | Major Freq                            | 0.53846                      | 0.65385                       | 0.5                           | 0.53846                       | 0.23077                       | 0.34615                       |
| n=13  | Minor Freq                            | 0.46154                      | 0.34615                       | 0.5                           | 0.46154                       | 0.76923                       | 0.65385                       |
| DOM   | Major Freq                            | 0.97619                      | 0.28571                       | 0.90476                       | 0.97368                       | 0.85714                       | 0.92857                       |
| n=21  | Minor Freq                            | 0.02381                      | 0.71429                       | 0.09524                       | 0.02632                       | 0.14286                       | 0.07143                       |
| COL   | Major Freq                            | 0.79412                      | 0.58491                       | 0.81132                       | 0.84314                       | 0.81132                       | 0.86792                       |
| n=53  | Minor Freq                            | 0.20588                      | 0.41509                       | 0.18868                       | 0.15686                       | 0.18868                       | 0.13208                       |
| CUB   | Major Freq                            | 0.95455                      | 0.24545                       | 0.96364                       | 0.98077                       | 0.85185                       | 0.94545                       |
| n=55  | Minor Freq                            | 0.04545                      | 0.75455                       | 0.03636                       | 0.01923                       | 0.14815                       | 0.05455                       |
| PUR   | Major Freq                            | 0.83333                      | 0.52778                       | 0.83333                       | 0.86111                       | 0.86111                       | 0.94444                       |
| n=18  | Minor Freq                            | 0.16667                      | 0.47222                       | 0.16667                       | 0.13889                       | 0.13889                       | 0.05556                       |
|       | <b>Loci<br/>Missing<br/>Maj./Min.</b> | <b>rs1462266</b><br>0<br>G/T | <b>rs17097005</b><br>0<br>T/G | <b>rs2869550</b><br>8<br>T/C  | <b>rs7198325</b><br>1<br>C/T  | <b>rs4470161</b><br>1<br>C/G  | <b>rs1019118</b><br>0<br>T/A  |
| ALL   | Major Freq                            | 0.32927                      | 0.94512                       | 0.88141                       | 0.95092                       | 0.78834                       | 0.58537                       |
| n=160 | Minor Freq                            | 0.67073                      | 0.05488                       | 0.11859                       | 0.04908                       | 0.21166                       | 0.41463                       |
| HUR   | Major Freq                            | 0.88462                      | 0.69231                       | 0.59091                       | 0.96154                       | 0.96154                       | 0.65385                       |
| n=13  | Minor Freq                            | 0.11538                      | 0.30769                       | 0.40909                       | 0.03846                       | 0.03846                       | 0.34615                       |
| DOM   | Major Freq                            | 0.26190                      | 0.97619                       | 0.88095                       | 0.75                          | 0.45238                       | 0.30952                       |
| n=21  | Minor Freq                            | 0.73810                      | 0.02381                       | 0.11905                       | 0.25                          | 0.54762                       | 0.69048                       |
| COL   | Major Freq                            | 0.34906                      | 0.92453                       | 0.81633                       | 1                             | 0.92308                       | 0.81132                       |
| n=53  | Minor Freq                            | 0.65094                      | 0.07547                       | 0.18367                       | 0                             | 0.07692                       | 0.18868                       |
| CUB   | Major Freq                            | 0.2                          | 0.99091                       | 0.98113                       | 0.98182                       | 0.77273                       | 0.52727                       |
| n=55  | Minor Freq                            | 0.8                          | 0.00909                       | 0.01887                       | 0.01818                       | 0.22727                       | 0.47273                       |
| PUR   | Major Freq                            | 0.38889                      | 1                             | 0.91667                       | 1                             | 0.80556                       | 0.5                           |
| n=18  | Minor Freq                            | 0.61111                      | 0                             | 0.08333                       | 0                             | 0.19444                       | 0.5                           |
|       | <b>Loci<br/>Missing</b>               | <b>rs17246021</b><br>0       | <b>rs12936629</b><br>0        | <b>rs3803828</b><br>0         | <b>rs6015771</b><br>0         | <b>rs1013001</b><br>0         | <b>rs2834567</b><br>0         |

|              | <b>Maj./Min.</b> | T/C               | T/G               | C/T              | C/A              | G/A              | A/G              |
|--------------|------------------|-------------------|-------------------|------------------|------------------|------------------|------------------|
| ALL<br>n=160 | Major Freq       | 0.89634           | 0.85061           | 0.97256          | 0.64024          | 0.63720          | 0.86585          |
|              | Minor Freq       | 0.10366           | 0.14939           | 0.02744          | 0.35976          | 0.36280          | 0.13415          |
| HUR<br>n=13  | Major Freq       | 0.61538           | 0.46154           | 0.80769          | 0.84615          | 0.23077          | 0.76923          |
|              | Minor Freq       | 0.38462           | 0.53846           | 0.19231          | 0.15385          | 0.76923          | 0.23077          |
| DOM<br>n=21  | Major Freq       | 0.95238           | 0.92857           | 1                | 0.5              | 0.76190          | 0.95238          |
|              | Minor Freq       | 0.04762           | 0.07143           | 0                | 0.5              | 0.23810          | 0.04762          |
|              | <b>Loci</b>      | <b>rs17246021</b> | <b>rs12936629</b> | <b>rs3803828</b> | <b>rs6015771</b> | <b>rs1013001</b> | <b>rs2834567</b> |
|              | <b>Missing</b>   | 0                 | 0                 | 0                | 0                | 0                | 0                |
|              | <b>Maj./Min.</b> | T/C               | T/G               | C/T              | C/A              | G/A              | A/G              |
| COL<br>n=53  | Major Freq       | 0.81132           | 0.79245           | 0.96226          | 0.84906          | 0.49057          | 0.70755          |
|              | Minor Freq       | 0.18868           | 0.20755           | 0.03774          | 0.15094          | 0.50943          | 0.29245          |
| CUB<br>n=55  | Major Freq       | 0.99091           | 0.91818           | 1                | 0.45455          | 0.78182          | 0.97273          |
|              | Minor Freq       | 0.00909           | 0.08182           | 0                | 0.54545          | 0.21818          | 0.02727          |
| PUR<br>n=18  | Major Freq       | 0.97222           | 0.97222           | 1                | 0.69444          | 0.75             | 0.94444          |
|              | Minor Freq       | 0.02778           | 0.02778           | 0                | 0.30556          | 0.25             | 0.05556          |
|              | <b>Loci</b>      | <b>rs440431</b>   | <b>rs1000472</b>  |                  |                  |                  |                  |
|              | <b>Missing</b>   | 1                 | 3                 |                  |                  |                  |                  |
|              | <b>Maj./Min.</b> | C/G               | T/C               |                  |                  |                  |                  |
| ALL<br>n=160 | Major Freq       | 0.59816           | 0.36957           |                  |                  |                  |                  |
|              | Minor Freq       | 0.40184           | 0.63043           |                  |                  |                  |                  |
| HUR<br>n=13  | Major Freq       | 0.77358           | 0.61538           |                  |                  |                  |                  |
|              | Minor Freq       | 0.22642           | 0.38462           |                  |                  |                  |                  |
| DOM<br>n=21  | Major Freq       | 0.60909           | 0.26364           |                  |                  |                  |                  |
|              | Minor Freq       | 0.39091           | 0.73636           |                  |                  |                  |                  |
| COL<br>n=53  | Major Freq       | 0.21429           | 0.21429           |                  |                  |                  |                  |
|              | Minor Freq       | 0.78571           | 0.78571           |                  |                  |                  |                  |
| CUB<br>n=55  | Major Freq       | 0.53846           | 0.29167           |                  |                  |                  |                  |
|              | Minor Freq       | 0.46154           | 0.70833           |                  |                  |                  |                  |
| PUR<br>n=18  | Major Freq       | 0.58333           | 0.26471           |                  |                  |                  |                  |
|              | Minor Freq       | 0.41667           | 0.73529           |                  |                  |                  |                  |

#### Supplemental Table S2.4: Allele Frequencies for the Setser80

Allele frequencies of the unrelated individuals from the GOAL dataset (n=160) are listed overall and per population. Missing refers to the number of genotypes missing from the full dataset.

Abbreviations used: Maj. = major allele, Min. = minor allele, Major Freq = major allele frequency, Minor Freq = minor allele frequency, HUR = Honduras, DOM = Dominican Republic, COL = Colombia, CUB = Cuba, and PUR = Puerto Rico.

**Supplemental Table S3.1: Country Attributable Mean  $F_{ST}$**

| SNP        | COL v.<br>CUB $F_{ST}$ | COL v.<br>DOM $F_{ST}$ | COL v.<br>HUR $F_{ST}$ | COL v.<br>PUR $F_{ST}$ | CUB vs.<br>DOM<br>$F_{ST}$ | CUB vs.<br>HUR $F_{ST}$ | CUB vs.<br>PUR $F_{ST}$ | DOM vs.<br>HUR $F_{ST}$ | DOM vs.<br>PUR $F_{ST}$    | HUR vs.<br>PUR $F_{ST}$    |
|------------|------------------------|------------------------|------------------------|------------------------|----------------------------|-------------------------|-------------------------|-------------------------|----------------------------|----------------------------|
| rs3777908  | 0.01129                | 0.00482                | 0.19754                | 0.00327                | 0.00583                    | 0.19560                 | 0.00425                 | 0.27318                 | 0.00207                    | 0.28808                    |
| rs9857908  | 0.12265                | 0.15319                | 0.21601                | 0.15047                | 0.20331                    | 0.10969                 | 0.07975                 | 0.15533                 | 0.09551                    | 0.07984                    |
| rs871234   | 0.07459                | 0.17146                | 0.09206                | 0.05378                | 0.15663                    | 0.01475                 | 0.00150                 | 0.24140                 | 0.24402                    | 0.03115                    |
| rs10981894 | 0.09794                | 0.07058                | 0.10767                | 0.13719                | 0.18782                    | 0.11750                 | 0.12768                 | 0.07918                 | 0.10264                    | 0.13199                    |
| rs16932385 | 0.02925                | 0.01002                | 0.26014                | 0.00765                | 0.06678                    | 0.22826                 | 0.06055                 | 0.30385                 | 0.02558                    | 0.31814                    |
| rs2051827  | 0.13258                | 0.14135                | 0.21359                | 0.14807                | 0.09160                    | 0.09166                 | 0.14127                 | 0.03493                 | 0.03900                    | 0.07688                    |
| rs178384   | 0.10567                | 0.07545                | 0.13410                | 0.10048                | 0.15544                    | 0.11984                 | 0.10358                 | 0.05628                 | 0.02126                    | 0.05578                    |
| rs7198325  | 0.09113                | 0.19367                | 0.03953                | 0.05931                | 0.18518                    | 0.01100                 | 0.02889                 | 0.27872                 | 0.25715                    | 0.00333                    |
| rs2834567  | 0.14717                | 0.15771                | 0.20497                | 0.16125                | 0.11549                    | 0.08042                 | 0.10691                 | 0.02358                 | 0.03044                    | 0.01768                    |
| rs1000472  | 0.16861                | 0.18993                | 0.20436                | 0.19454                | 0.08235                    | 0.04620                 | 0.05654                 | 0.03668                 | 0.04488                    | 0.00699                    |
| SNP        | Max $F_{ST}$           | Mean $F_{ST}$          | Min $F_{ST}$           | COL $F_{ST}$           | CUB $F_{ST}$               | DOM<br>$F_{ST}$         | HUR $F_{ST}$            | PUR $F_{ST}$            | 1 <sup>st</sup><br>Country | 2 <sup>nd</sup><br>Country |
| rs3777908  | 0.28808                | 0.09859                | 0.00021                | 0.05423                | 0.05424                    | 0.07148                 | 0.23860                 | 0.07442                 | HUR                        | PUR                        |
| rs9857908  | 0.21601                | 0.13657                | 0.00797                | 0.16058                | 0.12885                    | 0.15183                 | 0.14021                 | 0.10139                 | COL                        | DOM                        |
| rs871234   | 0.24402                | 0.10813                | 0.00015                | 0.09797                | 0.06186                    | 0.20338                 | 0.09484                 | 0.08261                 | DOM                        | COL                        |
| rs10981894 | 0.18782                | 0.11602                | 0.00706                | 0.10335                | 0.13274                    | 0.11006                 | 0.10909                 | 0.12488                 | CUB                        | PUR                        |
| rs16932385 | 0.31814                | 0.13102                | 0.00076                | 0.07676                | 0.09621                    | 0.10156                 | 0.27760                 | 0.10298                 | HUR                        | PUR                        |
| rs2051827  | 0.21359                | 0.11109                | 0.00349                | 0.15890                | 0.11428                    | 0.07672                 | 0.10426                 | 0.10131                 | COL                        | CUB                        |
| rs178384   | 0.15544                | 0.09279                | 0.00213                | 0.10392                | 0.12113                    | 0.07711                 | 0.09150                 | 0.07028                 | CUB                        | COL                        |
| rs7198325  | 0.27872                | 0.11479                | 0.00033                | 0.09591                | 0.07905                    | 0.22868                 | 0.08314                 | 0.08717                 | DOM                        | COL                        |
| rs2834567  | 0.20497                | 0.10456                | 0.00177                | 0.16778                | 0.11250                    | 0.08180                 | 0.08166                 | 0.07907                 | COL                        | CUB                        |
| rs1000472  | 0.20436                | 0.10311                | 0.00070                | 0.18936                | 0.08843                    | 0.08846                 | 0.07356                 | 0.07574                 | COL                        | DOM                        |

**Supplemental Table S3.1: Country Attributable Mean  $F_{ST}$** 

This table gives examples of how the  $F_{ST}$  calculations are made for each of the five countries. From left to right: section 1 gives the position information about that locus, section 2 presents the ten  $F_{ST}$  pairwise comparisons for each SNP, section 3 gives the standard maximum/mean/minimum, section 4 shows the final country attributable mean  $F_{ST}$  values for each country (the mean of the four pairwise comparisons with one country in common highlighted in bold in section 2), and section 5 identifies the highest (1<sup>st</sup>) and second highest (2<sup>nd</sup>) country attributable mean  $F_{ST}$  values and their corresponding country. Abbreviations used: HUR = Honduras, DOM = Dominican Republic, COL = Colombia, CUB = Cuba, and PUR = Puerto Rico.

**Supplemental Table S3.2: Description of the Setser80 Panel**

| <b>CHR</b> | <b>SNP</b> | <b>POS (hg18)</b> | <b>1<sup>st</sup> Country<br/>F<sub>ST</sub></b> | <b>1<sup>st</sup> Country</b> | <b>2<sup>nd</sup> Country<br/>F<sub>ST</sub></b> | <b>2<sup>nd</sup> Country</b> |
|------------|------------|-------------------|--------------------------------------------------|-------------------------------|--------------------------------------------------|-------------------------------|
| 1          | rs12130873 | 14629827          | 0.27834                                          | HUR                           | 0.08626                                          | PUR                           |
| 1          | rs6694897  | 25990886          | 0.16554                                          | COL                           | 0.11347                                          | CUB                           |
| 1          | rs1570099  | 35071891          | 0.14964                                          | COL                           | 0.09674                                          | CUB                           |
| 1          | rs10493701 | 81919788          | 0.21506                                          | HUR                           | 0.09218                                          | CUB                           |
| 1          | rs1040424  | 208201005         | 0.15556                                          | HUR                           | 0.09197                                          | CUB                           |
| 2          | rs10495889 | 41547938          | 0.16222                                          | HUR                           | 0.09780                                          | CUB                           |
| 2          | rs17018313 | 80045074          | 0.23693                                          | HUR                           | 0.07268                                          | PUR                           |
| 2          | rs7568419  | 177078765         | 0.22077                                          | HUR                           | 0.09977                                          | PUR                           |
| 2          | rs11693873 | 197850455         | 0.15274                                          | HUR                           | 0.09289                                          | CUB                           |
| 2          | rs4433950  | 239329179         | 0.23298                                          | HUR                           | 0.07530                                          | PUR                           |
| 3          | rs9853146  | 12274313          | 0.14598                                          | COL                           | 0.14106                                          | CUB                           |
| 3          | rs4685443  | 17205888          | 0.13890                                          | COL                           | 0.07537                                          | DOM                           |
| 3          | rs259425   | 22041732          | 0.16228                                          | COL                           | 0.10253                                          | DOM                           |
| 3          | rs2356298  | 51509761          | 0.16715                                          | HUR                           | 0.11188                                          | CUB                           |
| 3          | rs11719358 | 139267512         | 0.12963                                          | COL                           | 0.12198                                          | DOM                           |
| 3          | rs9857908  | 139318436         | 0.16058                                          | COL                           | 0.15183                                          | DOM                           |
| 3          | rs1586861  | 140541186         | 0.16976                                          | COL                           | 0.14602                                          | CUB                           |
| 4          | rs3910480  | 161880988         | 0.14157                                          | COL                           | 0.07363                                          | CUB                           |
| 5          | rs1366363  | 29985377          | 0.20338                                          | HUR                           | 0.11796                                          | CUB                           |
| 5          | rs3733838  | 42757753          | 0.21079                                          | HUR                           | 0.08908                                          | CUB                           |
| 5          | rs1438745  | 85249943          | 0.22267                                          | DOM                           | 0.09995                                          | PUR                           |
| 5          | rs16902270 | 85859474          | 0.16480                                          | COL                           | 0.10138                                          | CUB                           |
| 5          | rs871234   | 174418204         | 0.20338                                          | DOM                           | 0.09797                                          | COL                           |
| 5          | rs692713   | 176186041         | 0.20640                                          | DOM                           | 0.11848                                          | COL                           |
| 5          | rs4608884  | 178979749         | 0.14133                                          | COL                           | 0.10817                                          | CUB                           |
| 5          | rs190592   | 179263370         | 0.20994                                          | DOM                           | 0.13208                                          | COL                           |
| 6          | rs6596807  | 1267033           | 0.20170                                          | HUR                           | 0.09021                                          | CUB                           |
| 6          | rs9392285  | 1310265           | 0.19772                                          | HUR                           | 0.11633                                          | CUB                           |
| 6          | rs9501948  | 3156786           | 0.17955                                          | DOM                           | 0.11352                                          | COL                           |
| 6          | rs1329521  | 47840137          | 0.16820                                          | COL                           | 0.13484                                          | CUB                           |
| 6          | rs17745021 | 73769618          | 0.22153                                          | HUR                           | 0.08811                                          | PUR                           |
| 6          | rs3777908  | 111980345         | 0.23860                                          | HUR                           | 0.07442                                          | PUR                           |
| 6          | rs17087570 | 156949439         | 0.25097                                          | HUR                           | 0.11388                                          | CUB                           |
| 6          | rs4709836  | 164636380         | 0.14426                                          | COL                           | 0.10892                                          | DOM                           |
| 7          | rs12536738 | 95370320          | 0.18567                                          | HUR                           | 0.11446                                          | CUB                           |
| 7          | rs10953750 | 113424187         | 0.15649                                          | HUR                           | 0.09920                                          | CUB                           |
| 7          | rs2352479  | 137589866         | 0.12823                                          | COL                           | 0.09400                                          | DOM                           |
| 7          | rs17480133 | 145414794         | 0.19653                                          | COL                           | 0.11315                                          | CUB                           |
| 8          | rs766382   | 60170978          | 0.11474                                          | CUB                           | 0.11356                                          | COL                           |

| CHR | SNP        | POS (hg18) | 1 <sup>st</sup> Country<br>F <sub>ST</sub> | 1 <sup>st</sup> Country | 2 <sup>nd</sup> Country<br>F <sub>ST</sub> | 2 <sup>nd</sup> Country |
|-----|------------|------------|--------------------------------------------|-------------------------|--------------------------------------------|-------------------------|
| 8   | rs1588459  | 142143028  | 0.17997                                    | COL                     | 0.10768                                    | CUB                     |
| 9   | rs6474712  | 12405123   | 0.23846                                    | DOM                     | 0.08651                                    | COL                     |
| 9   | rs880397   | 12451349   | 0.22029                                    | DOM                     | 0.08143                                    | COL                     |
| 9   | rs10981894 | 115508956  | 0.13274                                    | CUB                     | 0.12488                                    | PUR                     |
| 10  | rs2008617  | 34499344   | 0.28372                                    | DOM                     | 0.10056                                    | PUR                     |
| 10  | rs16912280 | 59880729   | 0.26838                                    | DOM                     | 0.10208                                    | COL                     |
| 10  | rs1259603  | 76814933   | 0.23429                                    | HUR                     | 0.10085                                    | PUR                     |
| 10  | rs16932385 | 77017391   | 0.27760                                    | HUR                     | 0.10298                                    | PUR                     |
| 10  | rs11189628 | 100230671  | 0.15679                                    | COL                     | 0.10420                                    | CUB                     |
| 10  | rs17112705 | 101927184  | 0.17958                                    | COL                     | 0.11240                                    | CUB                     |
| 11  | rs1849352  | 61578905   | 0.18204                                    | HUR                     | 0.08943                                    | CUB                     |
| 11  | rs2878712  | 132343165  | 0.22463                                    | HUR                     | 0.09477                                    | PUR                     |
| 12  | rs10840730 | 17553352   | 0.22546                                    | HUR                     | 0.07914                                    | CUB                     |
| 12  | rs2051827  | 46242298   | 0.15890                                    | COL                     | 0.11428                                    | CUB                     |
| 12  | rs12146822 | 66857933   | 0.18173                                    | HUR                     | 0.15467                                    | CUB                     |
| 12  | rs7310083  | 68347053   | 0.21863                                    | PUR                     | 0.13093                                    | COL                     |
| 12  | rs1967232  | 115531283  | 0.14664                                    | COL                     | 0.10884                                    | CUB                     |
| 12  | rs6486527  | 129231989  | 0.11961                                    | COL                     | 0.10008                                    | DOM                     |
| 13  | rs9569702  | 56996265   | 0.15029                                    | HUR                     | 0.08848                                    | CUB                     |
| 13  | rs4341647  | 77872864   | 0.17141                                    | HUR                     | 0.09511                                    | CUB                     |
| 13  | rs9556940  | 97818960   | 0.16213                                    | COL                     | 0.10425                                    | CUB                     |
| 14  | rs1957572  | 67806224   | 0.17737                                    | HUR                     | 0.09433                                    | CUB                     |
| 14  | rs178384   | 79252594   | 0.12113                                    | CUB                     | 0.10392                                    | COL                     |
| 14  | rs12434466 | 96394042   | 0.20049                                    | HUR                     | 0.10147                                    | CUB                     |
| 14  | rs17094860 | 96560669   | 0.23314                                    | HUR                     | 0.10820                                    | CUB                     |
| 14  | rs12435621 | 97182294   | 0.31859                                    | HUR                     | 0.09892                                    | PUR                     |
| 14  | rs12431505 | 97188301   | 0.39018                                    | HUR                     | 0.12929                                    | PUR                     |
| 14  | rs1462266  | 97319155   | 0.24835                                    | HUR                     | 0.09924                                    | CUB                     |
| 14  | rs17097005 | 98013715   | 0.22045                                    | HUR                     | 0.08099                                    | PUR                     |
| 15  | rs2869550  | 76768056   | 0.15132                                    | HUR                     | 0.09375                                    | CUB                     |
| 16  | rs7198325  | 12572650   | 0.22868                                    | DOM                     | 0.09591                                    | COL                     |
| 16  | rs4470161  | 78627029   | 0.20305                                    | DOM                     | 0.12023                                    | COL                     |
| 17  | rs1019118  | 52103046   | 0.16663                                    | COL                     | 0.12606                                    | DOM                     |
| 17  | rs17246021 | 67616034   | 0.18769                                    | HUR                     | 0.11189                                    | CUB                     |
| 17  | rs12936629 | 67854083   | 0.22241                                    | HUR                     | 0.09781                                    | PUR                     |
| 20  | rs221308   | 34709812   | 0.16671                                    | COL                     | 0.10029                                    | DOM                     |
| 20  | rs6015771  | 58532301   | 0.16355                                    | COL                     | 0.14140                                    | CUB                     |
| 21  | rs1013001  | 14581125   | 0.16833                                    | HUR                     | 0.10737                                    | CUB                     |
| 21  | rs2834567  | 34950995   | 0.16778                                    | COL                     | 0.11250                                    | CUB                     |
| 21  | rs440431   | 42905781   | 0.18410                                    | DOM                     | 0.12326                                    | COL                     |

| CHR | SNP       | POS (hg18) | 1 <sup>st</sup> Country<br>F <sub>ST</sub> | 1 <sup>st</sup> Country | 2 <sup>nd</sup> Country<br>F <sub>ST</sub> | 2 <sup>nd</sup> Country |
|-----|-----------|------------|--------------------------------------------|-------------------------|--------------------------------------------|-------------------------|
| 21  | rs1000472 | 43602306   | 0.18936                                    | COL                     | 0.08846                                    | DOM                     |

**Supplemental Table S3.2: Description of the Setser80 Panel**

The Setser80 AIMs panel incorporates these 80 SNPs. Below appears the .map information (CHR = Chromosome, SNP = name of single nucleotide polymorphism, and POS = position in NCBI36/hg18 genome build). Mean F<sub>ST</sub> lists the average F<sub>ST</sub> across all 10 pairwise comparisons possible across 5 populations. The 1<sup>st</sup> country attributable mean F<sub>ST</sub> and 2<sup>nd</sup> country attributable mean F<sub>ST</sub> refers to the average of 4 pairwise comparisons that have one country in common. First country attributable mean F<sub>ST</sub> refers to the largest, most divergent average and its corresponding population while the 2<sup>nd</sup> country attributable mean F<sub>ST</sub> refers to the 2<sup>nd</sup> most divergent average and its population. Populations: HUR = Honduras, DOM = Dominican Republic, COL = Colombia, CUB = Cuba, and PUR = Puerto Rico.

**Supplemental Table S3.3: MLR Confusion Matrix**

| Known Origin | SNP Panel       | HUR | DOM | COL | CUB | PUR | PEL | MXL | Total |
|--------------|-----------------|-----|-----|-----|-----|-----|-----|-----|-------|
| <b>HUR</b>   | <b>Setser80</b> | 38  | 0   | 0   | 0   | 0   | 0   | 2   | 40    |
|              | <b>Seldin96</b> | 40  | 0   | 0   | 0   | 0   | 0   | 0   | 40    |
|              | <b>Kidd44</b>   | 23  | 0   | 6   | 0   | 3   | 0   | 8   | 40    |
| <b>DOM</b>   | <b>Setser80</b> | 0   | 39  | 0   | 1   | 0   | 0   | 0   | 40    |
|              | <b>Seldin96</b> | 0   | 33  | 0   | 3   | 4   | 0   | 0   | 40    |
|              | <b>Kidd44</b>   | 1   | 34  | 0   | 4   | 1   | 0   | 0   | 40    |
| <b>COL</b>   | <b>Setser80</b> | 0   | 0   | 31  | 0   | 2   | 0   | 7   | 40    |
|              | <b>Seldin96</b> | 1   | 0   | 33  | 1   | 4   | 0   | 1   | 40    |
|              | <b>Kidd44</b>   | 3   | 1   | 22  | 3   | 6   | 0   | 4   | 39*   |
| <b>CUB</b>   | <b>Setser80</b> | 0   | 0   | 0   | 40  | 0   | 0   | 0   | 40    |
|              | <b>Seldin96</b> | 0   | 4   | 0   | 34  | 1   | 0   | 0   | 39*   |
|              | <b>Kidd44</b>   | 0   | 4   | 1   | 29  | 6   | 0   | 0   | 40    |
| <b>PUR</b>   | <b>Setser80</b> | 0   | 0   | 0   | 3   | 37  | 0   | 0   | 40    |
|              | <b>Seldin96</b> | 0   | 3   | 5   | 5   | 27  | 0   | 0   | 40    |
|              | <b>Kidd44</b>   | 2   | 3   | 8   | 4   | 22  | 0   | 1   | 40    |
| <b>PEL</b>   | <b>Setser80</b> | 0   | 0   | 0   | 0   | 0   | 40  | 0   | 40    |
|              | <b>Seldin96</b> | 0   | 0   | 0   | 0   | 0   | 39  | 1   | 40    |
|              | <b>Kidd44</b>   | 0   | 0   | 0   | 0   | 0   | 37  | 3   | 40    |
| <b>MXL</b>   | <b>Setser80</b> | 0   | 0   | 4   | 0   | 0   | 0   | 36  | 40    |
|              | <b>Seldin96</b> | 0   | 0   | 0   | 0   | 0   | 0   | 40  | 40    |
|              | <b>Kidd44</b>   | 2   | 0   | 2   | 0   | 0   | 3   | 33  | 40    |

**Supplemental Table S3.3: MLR Confusion Matrix**

Confusion matrix showing into which population(s) each known population classifies.

This table reflects cumulative values from four sets of 70 micro-simulations (10 per population, per analysis) from the 7 Populations Combined dataset classified by MLR.

Abbreviations used: COL = Colombia, CUB = Cuba, DOM = Dominican Republic, HUR = Honduras, PUR = Puerto Rico, PEL = Peru from Lima, and MXL = Mexicans living in Los Angeles. \* One sample from this panel was unable to be classified for this population.

**Supplemental Table S3.4: Naïve Bayes Classification of Panels of 76 SNPs**

| SNP Panel                                | Dataset       | COL              | CUB              | DOM              | HUR              | PUR              | PEL              | MXL              | Overall |
|------------------------------------------|---------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|---------|
| <b>76 SNPs<br/>(removed<br/>1-3-5-7)</b> | <b>7 Pops</b> | 75.6%<br>(±6.2%) | 95.2%<br>(±2.5%) | 97.4%<br>(±1.5%) | 98.2%<br>(±0.8%) | 89.2%<br>(±5.0%) | 97.4%<br>(±1.3%) | 84.2%<br>(±2.3%) | 91.0%   |
| <b>Setser80</b>                          | <b>7 Pops</b> | 77.6%<br>(±8.2%) | 95.8%<br>(±1.9%) | 97.4%<br>(±1.7%) | 98.4%<br>(±0.9%) | 89.8%<br>(±2.9%) | 98%<br>(±1.0%)   | 83.4%<br>(±3.3%) | 91.5%   |
| <b>76 SNPs<br/>(removed<br/>2-4-6-8)</b> | <b>7 Pops</b> | 78.6%<br>(±5.2%) | 95%<br>(±1.6%)   | 96.6%<br>(±1.1%) | 97.8%<br>(±0.8%) | 88.8%<br>(±3.3%) | 98.4%<br>(±0.5%) | 82.6%<br>(±2.9%) | 91.1%   |
|                                          |               |                  |                  |                  |                  |                  |                  |                  |         |

**Supplemental Table S3.4: Naïve Bayes Classification of Panels of 76 SNPs**

Evaluation of the effect of four pairs of SNPs with  $LD\ r^2 > 0.5$  in the Setser80 on classification accuracy using the 7 Populations Combined dataset. One of each of the following pairs was removed for each subset of 76 SNPs where the number in parentheses corresponds to the SNPs removed: (1) rs11719358-rs9857908 (2), (3) rs6596807-rs9392285 (4), (5) rs1259603-rs16932385 (6), and (7) rs12435621-rs12431505 (8). Abbreviations used: 7 Pops = 7 Populations Combined, COL = Colombia, CUB = Cuba, DOM = Dominican Republic, HUR = Honduras, PUR = Puerto Rico, PEL = Peru from Lima, and MXL = Mexicans living in Los Angeles.

**Supplemental Table S3.5: MLR Classification of Panels of 76 SNPs**

| <b>SNP Panel</b>                 | <b>Dataset</b> | <b>COL</b>        | <b>CUB</b>       | <b>DOM</b>       | <b>HUR</b>       | <b>PUR</b>       | <b>PEL</b>       | <b>MXL</b>     | <b>Overall</b> |
|----------------------------------|----------------|-------------------|------------------|------------------|------------------|------------------|------------------|----------------|----------------|
| <b>76 SNPs (removed 1-3-5-7)</b> | <b>7 Pops</b>  | 77.5%<br>(±9.6%)  | 100%<br>(±0.0%)  | 97.5%<br>(±5.0%) | 92.5%<br>(±5.0%) | 92.5%<br>(±9.6%) | 100%<br>(±0.0%)  | 90%<br>(±0.0%) | <b>92.9%</b>   |
| <b>Setser80</b>                  | <b>7 Pops</b>  | 77.5%<br>(±9.6%)  | 100%<br>(±0.0%)  | 97.5%<br>(±5.0%) | 95%<br>(±5.8%)   | 92.5%<br>(±9.6%) | 100%<br>(±0.0%)  | 90%<br>(±8.2%) | <b>93.2%</b>   |
| <b>76 SNPs (removed 2-4-6-8)</b> | <b>7 Pops</b>  | 72.5%<br>(±12.6%) | 97.5%<br>(±5.0%) | 97.5%<br>(±5.0%) | 97.5%<br>(±5.0%) | 92.5%<br>(±9.6%) | 97.5%<br>(±5.0%) | 90%<br>(±8.2%) | <b>92.1%</b>   |

**Supplemental Table S3.5: MLR Classification of Panels of 76 SNPs**

Evaluation of the effect of four pairs of SNPs with  $LD\ r^2 > 0.5$  in the Setser80 on classification accuracy using the 7 Populations Combined dataset. One of each of the following pairs was removed for each subset of 76 SNPs where the number in parentheses corresponds to the SNPs removed: (1) rs11719358-rs9857908 (2), (3) rs6596807-rs9392285 (4), (5) rs1259603-rs16932385 (6), and (7) rs12435621-rs12431505 (8). Abbreviations used: 7 Pops = 7 Populations Combined, COL = Colombia, CUB = Cuba, DOM = Dominican Republic, HUR = Honduras, PUR = Puerto Rico, PEL = Peru from Lima, and MXL = Mexicans living in Los Angeles.

## BIBLIOGRAPHY

- The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. **467** (7319), 1061-1073 (2010 October 28).
- Abel, G. J. & Sander, N. Quantifying global international migration flows. *Science*. **343** (6178), 1520-1522 (2014 March 28).
- Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19** (9), 1655-1664 (2009 September).
- Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature*. **467** (7311), 52-58 (2010 September 2).
- Ambers, A. D. *et al.* Comprehensive forensic genetic marker analysis for accurate human remains identification using massively parallel DNA sequencing. *BMC Genomics*. **17** (Suppl 9), Article 750 (2016 October 17).
- Amirisetty, S., Hershey, G. K. K. & Baye, T. M. AncestrySNPminer: A bioinformatics tool to retrieve and develop ancestry informative SNP panels. *Genomics*. **100** (1), 57-63 (2012 July).
- Auton, A. *et al.* A global reference for human genetic variation. *Nature*. **526** (7571), 68–74 (2015 September 30).
- Ball, C. *et al.* Ethnicity Estimate White Paper. *ancestryDNA*. (2013 October 30).
- Bertoni, B., Budowle, B., Sans, M., Barton, S.A. & Chakraborty, R. Admixture in Hispanics: Distribution of ancestral population contributions in the continental United States. *Hum Bio*. **75** (1), 1-11 (2003 February).
- Bodner, M. *et al.* Rapid coastal spread of First Americans: Novel insights from South America's Southern Cone mitochondrial genomes. *Genome Res.* **22** (5), 811-820 (2012 May).
- Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *Am J Hum Genet*. **96** (1), 37-53 (2015 January 8).
- Buchmann, R.W. Genesis: Copyright © 2014, University of the Witwatersrand
- Burkart, K. M. *et al.* A genome-wide association study in Hispanics/Latinos identifies novel signals for lung function – The Hispanic Community Health Study/Study of Latinos. *Am J Resp Crit Care Med*. **198** (2), 208–219 (2018 July 15).
- Butler, J. M. (2<sup>nd</sup> ed.) Forensic DNA typing: Biology, technology, and genetics of STR markers. (Elsevier, 2005).

Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L. & Nickerson, D. A. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet.* **74** (1), 106-120 (2004 January).

Cann, H. M. *et al.* A human genome diversity cell line panel. *Science.* **296** (5566), 261-262 (2002 April 12).

Cheung, E. Y. Y., Gahan, M. E. & McNevin, D. Prediction of biogeographical ancestry from genotype: A comparison of classifiers. *Int J Legal Med.* **131** (4), 901-912 (2017 July 1).

Cockerham, C. C. & Weir, B. S. Estimation of gene flow from F-Statistics. *Evolution.* **47** (3), 855-863 (1993 June).

Conomos, M. P. *et al.* Genetic diversity and association studies in US Hispanic/Latino populations: Applications in the Hispanic Community Health Study/Study of Latinos. *Am J Hum Genet.* **98** (1), 165-184 (2016 January 7).

Das, R. & Upadhyai, P. An ancestry informative marker set which recapitulates the known fine structure of populations in South Asia. *Genome Biol Evol.* **10** (9), 2408-2416 (2018 September 1).

De la Puente, M. *et al.* The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs. *Forensic Sci Int-Gen.* **22**, 81-88 (2016 May).

Ding, L. *et al.* Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC Genomics.* **12**, 622; 10.1186/1471-2164-12-622 (2011 December 20).

Durand, E. Y., Do, C. B., Mountain, J. L. & Macpherson, J. M. Ancestry composition: A novel, efficient pipeline for ancestry deconvolution. 23andMe White Paper 23-16. (2014 October 17).

Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour.* **4** (2), 359-361 (2012).

Eduardoff, M. *et al.* Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM™. *Forensic Sci Int-Gen.* **23**, 178-189 (2016 July 1).

Elhaik, E. *et al.* The GenoChip: A new tool for genetic anthropology. *Genome Biol Evol.* **5** (5), 1021-1031 (2013).

- Elhaik, E. *et al.* Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat Commun.* **5**, 3513; 10.1038/ncomms4513 (2014 April 29).
- Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol Ecol.* **14** (8), 2611-2620 (2005 July).
- Fondevila, M. *et al.* Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies. *Forensic Sci Int-Gen.* **7** (1), 63-74 (2013 January).
- Fortes-Lima, C. *et al.* Exploring Cuba's population structure and demographic history using genome-wide data. *Sci Rep.* **8** (1), 11422; 10.1038/s41598-018-29851-3 (2018 December 1).
- Francioli, L. & MacArthur, D. gnomAD v3.0. MacArthur Lab Blog. (2019 October 16). URL <https://macarthurlab.org/blog/>
- Freire-Aradas, A. *et al.* Exploring iris colour prediction and ancestry inference in admixed populations of South America. *Forensic Sci Int-Gen.* **13**, 3-9 (2014 November).
- Frey, W. H. Census projects new “majority minority” tipping points. *Brookings.* (2012 December 13). URL <https://www.brookings.edu/opinions/census-projects-new-majority-minority-tipping-points/>
- Galanter, J. M. *et al.* Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genet.* **8** (3), e1002554; 10.1371/journal.pgen.1002554 (2012 March).
- Gao, C. *et al.* A comprehensive analysis of common and rare variants to identify adiposity loci in Hispanic Americans: The IRAS family study (IRASFS). *PLoS ONE.* **10** (11) e0134649; 10.1371/journal.pone.0134649 (2015 November 1).
- Genetic Information Nondiscrimination Act of 2008, Pub. L. No. 110-233 (2008). <https://www.eeoc.gov/laws/statutes/gina.cfm>
- Granados-Silvestre, M. A. *et al.* Susceptibility background for type 2 diabetes in eleven Mexican Indigenous populations: HNF4A gene analysis. *Mol Genet Genomics.* **292** (6), 1209-1219 (2017 December 1).
- Gurdasani, D. *et al.* The African genome variation project shapes medical genetics in Africa. *Nature.* **517** (7534), 327-332 (2015 January 15).
- Hartl, D. L. & Clark, A. G. (3<sup>rd</sup> ed.) Principles of Population Genetics. (Sinauer Associates, Inc., 1997).

Heinz, T. *et al.* Ancestry analysis reveals a predominant Native American component with moderate European admixture in Bolivians. *Forensic Sci Int-Gen.* **7** (5), 537-542 (2013).

Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science.* **343** (6172), 747-751, (2014 February 14).

Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating, and interpreting  $F_{ST}$ . *Nat Rev Genet.* **10** (9), 639-650 (2009 September).

Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5** (6), e1000529; 10.1371/journal.pgen.1000529 (2009 June).

Howie, B. & Marchini, J. Instructions for IMPUTE version 2. (2009 June 18). URL [https://mathgen.stats.ox.ac.uk/impute/impute\\_v2\\_instructions.pdf](https://mathgen.stats.ox.ac.uk/impute/impute_v2_instructions.pdf)

Huerta-Chagoya, A. *et al.* A panel of 32 AIMs suitable for population stratification correction and global ancestry estimation in Mexican mestizos. *BMC Genetics.* **20** (1), 5; 10.1186/s12863-018-0707-7 (2019 January 8).

Infinium Multi-Ethnic AMR/AFR BeadChip data sheet. Illumina. Pub. No. 370-2015-006. (2016 February 29).

The International HapMap Consortium. The International HapMap Project. *Nature.* **426** (6968), 789-796 (2003 December 18).

Jakobsson, M. & Rosenberg, N. A. CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics.* **23** (14), 1801-1806 (2007 July 15).

Jakobsson, M., Edge, M. D. & Rosenberg, N. A. The relationship between  $F_{ST}$  and the frequency of the most frequent allele. *Genetics.* **193** (2), 515-528 (2013 February).

Jeong, D. H., Ziemkiewicz, C., Ribarsky, W. & Chang, R. Understanding principal component analysis using a visual analytics tool. Charlotte Visualization Center. (UNC Charlotte, 2008).

Jia, J., Wei, Y., Qin, C., Hu, L., Wan, L. & Li, C. Developing a novel panel of genome-wide ancestry informative markers for bio-geographical ancestry estimates. *Forensic Sci Int-Gen.* **8**, 187-194 (2014).

- Jin, X.-Y. *et al.* A set of novel SNP loci for differentiating continental populations and three Chinese populations. *PeerJ.* **7**, e6508; 10.7717/peerj.6508 (2019 March 29).
- Johnston, H. R. *et al.* Identifying tagging SNPs for African specific genetic variation from the African Diaspora Genome. *Sci Rep.* **7**, 46398; 10.1038/srep46398 (2017 April 21).
- Kent, W. J. *et al.* UCSC Genome Browser: The human genome browser at UCSC. *Genome Res.* **12** (6), 996-1006 (2002 June). URL <http://genome.ucsc.edu>.
- Khan, R. Basic concepts - linkage disequilibrium. *Gene Expression.* (2007 January 24). URL <https://www.gnXP.com/WordPress/2007/01/24/basic-concepts-linkage-disequilibrium/>
- Kidd, J., Friedlaender, F. R., Speed, W. C., Pakstis, A. J., De La Vega, F. M. & Kidd, K. K. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investig Genet.* **2** (1), 1; 10.1186/2041-2223-2-1 (2011 January 5).
- Kidd, K. K., Kidd, J. R., Pakstis, A. J., Speed, W. C. & Donnelly, M. P. Developing SNP Panels for ancestry identification useful in forensic investigations. Poster. (2011).
- Kidd, K. K. *et al.* Microhaplotype loci are a powerful new type of forensic marker. *Forensic Sci Int-Gen.* Supp Series 4 (1), e123–e124; 10.1016/j.fsigs.2013.10.063 (2013).
- Kidd, K. K. *et al.* Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci Int-Gen.* **10** (1), 23-32 (2014 May).
- Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature.* **488** (7412), 471-475 (2012 August 23).
- Kosoy, R. *et al.* Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat.* **30** (1), 69-78 (2009 January).
- Kusev, P., van Schaik, P., Tsaneva-Atanasova, K., Juliusson, A. & Chater, N. Adaptive anchoring model: How static and dynamic presentations of time series influence judgment predictions. *Cogn Sci.* **42** (1), 77-102 (2018 January).
- Lai, C.-Q. *et al.* Population admixture associated with disease prevalence in the Boston Puerto Rican health study. *Hum Genet.* **125** (2), 199-209 (2009).
- Lalueza-Fox, C., Gilbert, M. T. P., Martínez-Fuentes, A. J., Calafell, F. & Bertranpetit, J. Mitochondrial DNA from Pre-Columbian Ciboneys from Cuba and

the prehistoric colonization of the Caribbean. *Am J of Phys Anthro.* **121** (2), 97-108 (2003 June 1).

Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature.* **513** (7518), 409-413 (2014 September 18).

Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature.* **519** (7543), 309-314 (2015 March 19).

Lorente, J. A. Trafficking in human beings: modern slavery. EndSlavery. Workshop 2-3, November 2013. URL [http://www.endslavery.va/content/endslavery/en/publications/scripta\\_varia\\_122/lorente.html](http://www.endslavery.va/content/endslavery/en/publications/scripta_varia_122/lorente.html) (2019).

MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Res.* **45** (D1), D896-D901 (2017 January 1).

Machiela, M. J. & Chanock, S. J. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics.* **31** (21), 3555-3557 (2015 December 18). URL <https://ldlink.nci.nih.gov>

Manichaikul, A., *et al.* Population structure of Hispanics in the United States: The multi-ethnic study of atherosclerosis. *PLoS Genet.* **8** (4), e1002640; 10.1371/journal.pgen.1002640 (2012 April 12).

Marcheco-Teruel, B. *et al.* Cuba: Exploring the history of admixture and genetic basis of pigmentation using autosomal and uniparental markers. *PLoS Genet.* **10** (7), e1004488; 10.1371/journal.pgen.1004488 (2014 July 14).

Maroñas, O. *et al.* Development of a forensic skin colour predictive test. *Forensic Sci Int-Gen.* **13**, 34-44 (2014 November).

McDonald, J. H. (3<sup>rd</sup> ed.) Multiple Regression, 229-237. Handbook of Biological Statistics. (Sparky House Publishing, 2014). URL <http://www.biostathandbook.com/multipleregression.html>

McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet.* **40** (10), 1166-1174 (2008 October).

McNevin, D. *et al.* An assessment of Bayesian and multinomial logistic regression classification systems to analyse admixed individuals. *Forensic Sci Int-Gen. Supplement Series* **4** (1), e63-e64; 10.1016/j.fsigs.2013.10.032 (2013).

- Moreno-Estrada, A. *et al.* Reconstructing the population genetic history of the Caribbean. *PLoS Genet.* **9** (11), e1003925; 10.1371/journal.pgen.1003925 (2013 November 14).
- Moreno-Estrada, A. *et al.* The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science.* **344** (6189), 1280-1285 (2014 June 13).
- National Cancer Institute, Division of Cancer Epidemiology & Genetics. LD Matrix. URL <https://ldlink.nci.nih.gov/?tab=ldmatrix> (2019).
- National DNA Index System (NDIS) Operations Manual, version 8. FBI Laboratory. (2019 May 1).
- Negroponte, D. V. The surge in unaccompanied children from Central America: A humanitarian crisis at our border. *Brookings.* (2014 July 2). URL <https://www.brookings.edu/blog/up-front/2014/07/02/the-surge-in-unacc...d-children-from-central-america-a-humanitarian-crisis-at-our-border/>
- Nelson, M. R. *et al.* The Population Reference Sample, POPRES: A Resource for population, disease, and pharmacological genetics research. *Am J Hum Genet.* **83** (3), 347-358 (2008 September 12).
- Nievergelt, C. M. *et al.* Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investig Genet.* **4** (1), Article 13 (2013 July 13).
- Norris, E. T. *et al.* Genetic ancestry, admixture and health determinants in Latin America. *BMC Genomics.* **19** (Suppl 8), Article 861 (2018 December).
- Pabon-Nau, L. P., Cohen, A., Meigs, J. B. & Grant, R. W. Hypertension and diabetes prevalence among U.S. Hispanics by country of origin: The National Health Interview Survey 2000-2005. *J Gen Intern Med.* **25** (8), 847-852 (2010 August).
- Paschou, P., Lewis, J., Javed, A. & Drineas, P. Ancestry informative markers for fine-scale individual assignment to worldwide populations. *J Med Genet.* **47** (12), 835-847 (2010 December).
- Patterson, N. *et al.* Methods for high-density admixture mapping of disease genes. *Am J Hum Genet.* **74** (5), 979-1000 (2004 May).
- Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2** (12), 2074-2093; 10.1371/journal.pgen.0020190 (2006 December 22).

- Pereira, R. *et al.* Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. *PLoS ONE*. **7** (1), e29684; 10.1371/journal.pone.0029684 (2012 January 17).
- Phillips, C. *et al.* Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int-Gen*. **1** (3-4), 273-280 (2007 December).
- Phillips, C. *et al.* Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set. *Forensic Sci Int-Gen*. **11** (1), 13-25 (2014 July).
- Phillips C. Forensic genetic analysis of biogeographic ancestry. *Forensic Sci Int-Gen*. **18**, 49-65 (2015).
- Pośpiech, E. *et al.* The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction. *Forensic Sci Int-Gen*. **11**, 64-72 (2014 July).
- Price, A. L. *et al.* A genomewide admixture map for Latino populations. *Am J Hum Genet*. **80** (6), 1024-1036 (2007 June).
- Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics*. **155** (2), 945-959 (2000 June).
- Pritchard, J. K., Wen, X. & Falush, D. Documentation for STRUCTURE software: Version 2.3. (2010 February 2).
- Purcell, S. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. **81** (3), 559-575 (2007 September).
- Purcell, S. & Chang, C. PLINK 1.9. URL <http://www.cog-genomics.org/plink/1.9/>.
- Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*. **505** (7481), 87-91 (2014 January 2).
- Raghavan, M. *et al.* Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science*. **349** (6250), aab3884; 10.1126/science.aab3884 (2015 August 21).
- Rajeevan, H., Soundararajan, U., Kidd, J., R., Pakstis, A. & Kidd, K., K. ALFRED: An allele frequency resource for research and teaching. *Nucleic Acids Res*. **40** (D1), D1010-D1015 (2012 January).
- Rasmussen, M. *et al.* The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature*. **506** (7487), 225-229 (2014 February 13).

Reich, D. *et al.* Linkage disequilibrium in the human genome. *Nature*. **411** (6834), 199-204 (2001 May 10).

Reich, D. *et al.* Reconstructing Native American population history. *Nature*. **488** (7411), 370-374 (2012 August 16).

Rite Aid HomeDNA Paternity Test, 2007. URL <https://www.riteaid.com/shop/homedna-paternity-test-for-at-home-use-1-ct-8022392>.

Ritter, N. Missing persons and unidentified remains: The nation's silent mass disaster. *NIJ Journal*. **256** (2007).

Rosenberg, N. A. Distruct: A program for the graphical display of population structure. *Mol Ecol Notes*. **4** (1), 137-138 (2004 March).

Ruiz, Y. *et al.* Further development of forensic eye color predictive tests. *Forensic Sci Int-Gen*. **7** (1), 28-40 (2013 January).

Ruiz-Linares, A. *et al.* Admixture in Latin America: Geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet*. **10** (9), e1004572; 10.1371/journal.pgen.1004572 (2014 September 1).

Russell, J. G. Gedmatch: a DNA geek's dream site. *The Legal Genealogist*. (2012 August 12). URL <https://www.legalgenealogist.com/2012/08/12/gedmatch-a-dna-geeks-dream-site/>

Russell, J. G. GEDmatch reverses course. *The Legal Genealogist*. (2019 May 19). URL <https://www.legalgenealogist.com/2019/05/19/gedmatch-reverses-course/>

Santos, C. *et al.* Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific Region. *Forensic Sci Int-Gen*. **20**, 71-80 (2016 January 1).

Santos, H. C. *et al.* A minimum set of ancestry informative markers for determining admixture proportions in a mixed American population: the Brazilian set. *Eur J Hum Genet*. **24** (5), 725-731 (2016 May 1).

Santangelo, R., González-Andrade, F., Børsting, C., Torroni, A., Pereira, V. & Morling, N. Analysis of ancestry informative markers in three main ethnic groups from Ecuador supports a trihybrid origin of Ecuadorians. *Forensic Sci Int-Gen*. **31**, 29-33 (2017 November).

Schroeder, H. *et al.* Genome-wide ancestry of 17th-century enslaved Africans from the Caribbean. *PNAS*. **112** (12), 3669-3673 (2015 March 24). [www.slavevoyages.org](http://www.slavevoyages.org)

- Shriver, M. *et al.* The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum Genomics*. **1** (4), 274-286 (2004 May).
- Singh, S. Understanding the bias-variance tradeoff. (2018 May 20). URL <https://towardsdatascience.com/understanding>
- Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell*. **177** (1), 26-31 (2019 March 21).
- Skoglund, P. *et al.* Genetic evidence for two founding populations of the Americas. *Nature*. **525** (7567), 104-108 (2015 September 3).
- Söchtig, J. *et al.* Exploration of SNP variants affecting hair colour prediction in Europeans. *Int J Legal Med*. **129** (5), 963-975 (2015 September).
- Soundararajan, U., Yun, L., Shi, M. & Kidd, K. K. Minimal SNP overlap among multiple panels of ancestry informative markers argues for more international collaboration. *Forensic Sci Int-Gen*. **23**, 25-32 (2016 July 1).
- Taboada-Echalar, P. *et al.* The genetic legacy of the pre-colonial period in contemporary Bolivians. *PLoS ONE*. **8** (3), e58980; 10.1371/journal.pone.0058980 (2013 March).
- Tamm, E., *et al.* Beringian standstill and spread of Native American founders. *PLoS ONE*. **2** (9), e829; 10.1371/journal.pone.0000829 (2007 September 5).
- Tandy-Connor, S. *et al.* False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care. *Genet in Med*. **20** (12), 1515-1521 (2018 December 1).
- Tishkoff, S. A. & Kidd, K. K. Implications of biogeography of human populations for 'race' and medicine. *Nature Genet*. **36** (11S), S21-S27; 10.1038/ng1438 (2004 November).
- Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science*. **324** (5930), 1035-1044 (2009 May 22).
- Tian, C. *et al.* Analysis of East Asia genetic substructure using genome-wide SNP arrays. *PLoS ONE*. **3** (12), e3862; 10.1371/journal.pone.0003862 (2008 December 5).
- Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science*. **343** (6174), 1017-1021 (2014 February 28).

Wathen, M. J., Gautam, Y, Ghandikota, S., Rao, M. B. & Marsha, T. B. LEI: A novel allele frequency-based Feature Selection method for multi-ancestry admixed populations. *Sci Rep.* **9** (1), e11103; 10. 1038/s41598-019-47012-y (2019 July 31).

Wee, S. China uses DNA to track its people, with the help of American expertise. *New York Times.* (2019 February 21).

Weir, B. S. & Cockerham, C. C. Estimation of gene flow from F-statistics. *Evolution.* **47** (3), 855-863 (1993).

Wright, S. The genetical structure of populations. *Ann Eugenica.* **15** (4), 323-354 (1951 March).

Yahya, P. *et al.* Analysis of the genetic structure of the Malay population: Ancestry-informative marker SNPs in the Malay of Peninsular Malaysia. *Forensic Sci Int-Gen.* **30**, 152–159 (2017 September).

Yuan, X., Miller, D. J., Zhang, J., Herrington, D. & Wang, Y. An overview of population genetic data simulation. *J Comput Biol.* **19** (1), 42-54 (2012 January 1).

Yudell, M., Roberts, D., DeSalle, R. & Tishkoff, S. Taking race out of human genetics. *Science.* **351** (6273), 564-565 (2016 February 5).

Zeng, X., Chakraborty, R., King, J. L., LaRue, B., Moura-Neto, R. S. & Budowle, B. Selection of highly informative SNP markers for population affiliation of major US populations. *Int J Legal Med.* **130** (2), 341–352 (2016 March).

Zeng, X. Selection of highly informative markers for apportionment of ancestry and population affiliation. Fort Worth, TX: University of North Texas Health Science Center. (2016 May 1).