Chapter 3

Identification of recent admixture in an Indian population of African ancestry
3.1. Introduction

Enigmatic history of India has been shaped by different waves of migration and admixture events [1-5]. Admixed populations not only provides insights about migration of populations from different ancestries but are also a potential resource for mapping disease loci [6-8] and detecting signatures of selection [9-11]. An admixture event between populations leads to an extended linkage disequilibrium (LD), which could greatly facilitate the mapping of human disease loci [12]. The power of gene mapping by admixture linkage disequilibrium (MALD) primarily depends upon prevalence of a disease/trait among ancestral populations and extent of LD [8, 12, 13]. Contrasting phenotypes or traits of ancestral populations and history of recent admixture increase the power of study [14]. Admixed populations like African Americans and Latinos in the United States have been largely studied for admixture mapping [6-8, 15]. Admixture mapping has proven successful for identifying genes implicated in diseases like prostate cancer [15], Type 2 diabetes [16, 17], end-stage kidney diseases [18] etc. Recently, admixture events in Asian populations like the Uyghurs have also been reported [19, 20]. Reich et al. have proposed that ancestry of Indian populations has majorly comes from two groups – ANI (Ancestral North Indians) and ASI (Ancestral South Indians) and populations of India have arisen from ancient admixture events of these diverged groups, because of which extent of LD in these admixed populations is small [5]. Furthermore, in populations within India the difference in allele frequency in the ancestral populations is small and hence might not be useful for MALD [5]. In this study, I take a detailed look at a population which is known to be an out-group population from the Indian Genome Variation Consortium (IGVC) and many other supporting studies. This population, known as Siddi, has been given a nomenclature of OG-W-IP1 by convention of IGVC as it is an out-group (OG) isolated population (IP) from the western (W) part of India. OG, reckoned to be the “lost tribe of Africa” is one of the major nonnative tribal communities of Gujarat and they have adapted to the language and the religious practices of the place.

It is reported that African slaves are the ancestors of this tribal community and they came to India during the 12-15th century with the Arab merchants [21]. It is also believed that the Portuguese merchants brought the African slaves to the west coast of India, possibly in Karnataka and Maharashtra. They eventually expanded and migrated northward. Apart from Gujarat, this community resides in some parts of Karnataka, Goa and Maharashtra.
Interestingly, the area in Gujarat, where this tribe resides is extremely saline and most of the salt that is exported out of India is produced in this area. The anthropological and other evidences linking this OG population to their African ancestry has been weak and limited primarily to musical instruments, folklores and traditions [22]. There were few studies [2, 23] that attempted to decipher the genetic ancestry of this population.

In this chapter, I demonstrate that the OG population derives its ancestry from both Africa and India. Allelic differences between important genic SNPs in ancestral populations make this population a potential candidate for admixture mapping. Comparison of LD data revealed that OG has higher LD in comparison to both its ancestral populations, which is an indication for recent population admixture. As indigenous and migrant populations from two different continents, inter-mated and subsequently formed the admixed population, it provided novel opportunities for natural selection to occur. Functional annotation clustering of markers that have differential allele frequency in the two ancestral populations revealed significant enrichment of ion channel genes especially related to potassium transport and cadherins. This study not only provides a window to look into their past but also evaluated them as a resourceful population for disease gene mapping.

### 3.2. Methodology

#### 3.2.1. Population data sets

Three population datasets were used for analysis. (i) A subset of 26 reference populations of IGVC [2] comprising of 509 samples from diverse linguistic groups residing in different geographical regions. These include Austro-Asiatic (AA), Tibeto-Burman (TB), Dravidian (DR), Indo-European (IE) linguistic origins from North (N), East (E), West (W), South (S), North-East (NE) and central (C) part of India and one out-group population of African origin (OG). The details of population identification, sample collection and DNA isolation are described elsewhere [2]. For naming of populations we have followed a convention where each population was represented by their linguistic affiliation followed by geographical location and ethnicity (*Supplementary Table 3.A*). (ii) Second set comprised of 210 samples from 4 populations of the International HapMap Project [24] [(60 CEU (Utah residents with ancestry from northern and western

* Supplementary data provided in CD.
Europe), 60 YRI (Yoruba in Ibadan, Nigeria), 45 CHB (Han Chinese in Beijing) and 45 JPT (Japanese in Tokyo)], (iii) Third dataset included 52 populations comprising 1043 CEPH-HGDP Samples (Centre d’E’tude du Polymorphisme Humain [CEPH]-obtained samples from the Human Genome Diversity Panel [HGDP]) [25].

### 3.2.2. Genotype data sets

Genotype data on the above mentioned was used for the analysis. (i) 509 IGVC samples generated using Affymetrix 50k Xba1 240 GeneChip Human Mapping array (Affymetrix, Santa Clara, CA USA) as a part of the IGVC project [26], (ii) genotypes on 210 samples of International HapMap Project [24] and (iii) 1043 samples of HGDP-CEPH Human Genome Diversity Panel generated on Illumina Human Hap650K Beadchip [25]. Genotypes of common set of 18,534 SNPs from the three datasets (IGVC, HapMap and HGDP-CEPH) were merged and used for further analysis. It was ensured that these SNPs were from same strand and pass all the standard QC criteria. The SNPs that showed deviation from Hardy-Weinberg equilibrium within population were excluded from data analysis. The physical positions of the SNPs were retrieved from Homo sapiens NCBI build 36. The average spacing between adjacent markers is 166.7kb with minimum and maximum distance of 17 bp and 31.5 Mb respectively. An overview of study design is described in **Figure 3.1**.

![Figure 3.1: Overall design of the study for elucidating the genomic architecture of an admixed population.](image)
3.2.3. Population genetic analysis

Principal Component Analysis (PCA) was performed using EIGENSOFT 3.0 [27, 28]. Model based clustering algorithm, STRUCTURE [29-31] was used for estimation of individual and population ancestries. STRUCTURE analysis was performed assuming two and three clusters (K=2, 3) with 20,000 burnin period and 20,000 iterations. ADMIXTURE [32] software was also used to estimate individual ancestry proportions and validate our STRUCTURE results. Analysis of Molecular Variance (AMOVA) was calculated as per Excoffier et al., [33] using the software package ARLEQUIN [34]. $F_{ST}$ and Reynold's distance was computed to estimate the extent of genetic differentiation between populations. $F_{ST}$ and its significance were calculated for all 18,534 markers using Weir and Hill method [35] in ARLEQUIN with 10000 permutations which adjusts for sample size variation across populations. PLINK was used to estimate pairwise LD ($r^2$) [36] for all the SNPs on one chromosome separately for one population at a time. The average LD per 200000 bases was plotted.

3.2.4. Functional annotation of ancestry informative markers

A set of ancestry informative markers (AIMs) were defined for the ancestral populations of OG using complete set of 18,534 SNPs. Allelic frequency of these markers differ substantially between the ancestral populations. Two hypothetical populations which can serve as putative ancestors to OG were defined (details in the next section). The putative African ancestry comes from a hypothetical population consisting of 32 individuals (11 belonging to Bantu (Kenya), 21 Yoruba) and the non-African ancestor population consisting 86 individuals from four different IE speaking groups (24 from IE-N-LP10, 23 from IE-N-LP18, 20 from IE-W-LP2, 19 from IE-W-LP4).

There were 3396 SNPs that have an $F_{ST}$ value $> 0.1$ between the two ancestral populations (Supplementary Table 3.B). These ancestry informative markers (AIMs) were mapped to genes reported to be associated to diseases in Genetic Association Database (GAD) using the SNP NEXUS tool [37]. Whether OG had specific functional enrichment of genes which can be attributable to either of their ancestors has also been explored. For this, closeness of OG to either of the ancestors was computed by comparing the allele frequency of AIMs in OG with the Indian and African ancestral populations. All those AIMs from analysis whose frequencies were similar to their expected frequency were excluded, i.e. within a cutoff of 5% of the weighted average (weights used are the
approximate ancestry estimates of 0.59 for African and 0.41 from Indian populations) of the ancestral allele frequencies. The AIMs were then binned into two groups, one ‘close’ to African ancestors in terms of allele frequency and one close to the Indian ancestor’s. A web based tool DAVID [38] was used for their functional gene classification as well as functional annotation clustering. For classification of genes highest stringency criteria was used and a cutoff $> 1.5$ was set. Functional Annotation clustering was also carried out at highest stringency. The results with $> 3$ fold enrichments at $\leq 1\%$ FDR have been represented (Table 3.5 and Supplementary Table 3.C).

3.3. Results

3.3.1. PCA analysis for identification of putative ancestor populations to the OG

Initial Principal Component Analysis (PCA) with 26 IGV populations suggested that OG population was distant from all TB and IE isolated populations of North and North Eastern regions as well as Austro-Asiatic and Dravidian isolated populations of IGV along both the principal components (Figure 3.2).

**Figure 3.2:** Principal component analysis of 26 Indian populations showing the Siddis as an outgroup (OG). The second eigenvector explains the separation of the OG population from other Indian populations. The first eigenvector explains the variation in rest 25 population groups. Along the first eigenvector, the populations on the left are primarily Tibeto-Burman (TB) speaking population who separate from the Indo-European (IE) and Dravidian (DR) speaking populations. The populations are coded by linguistic lineage (AA, Austro-Asiatic; IE, Indo-European; DR, Dravidian and TB, Tibeto-Burman) followed by geographical location (N, north; NE, north-east; W, west; E, east; S, south and C, central) and ethnic category (LP, castes; SP, religious groups and IP, tribes).
All distant IGV populations were excluded from analysis and carried out next level of PCA with the remaining 18 IGV populations including OG. In the search of the possible African ancestor(s) to OG, all African populations from the HGDP–CEPH panel were included excluding the Mozambites who are highly admixed between Africans and Middle Easterners [25]. HGDP populations from Pakistan were also included because of their geographical proximity to OG. As history states that the OG was brought into India by Portuguese traders [39], CEU population from HapMap was included. This integrated analysis was performed on common set of 18534 markers in IGV, HGDP–CEPH and HapMap populations.

The PCA along the first principal component (PC1) separated all African from the non-African populations and explained 6.5% of the entire variation; separation of the various non-African groups was observed along PC2 (Figure 3.3).

![Figure 3.3: Genetic relatedness of OG with populations of Indian subcontinent and Africa. Principal Component Analysis of 17 Indian (IE speakers and DR speakers) populations, 7 African and 4 Pakistani populations from HGDP, CEU from HapMap and the OG clearly shows that the OGs are admixed between Indians and Africans and there is no contribution of CEU.](image)

Noteworthy, the separation of the non-African population groups are achieved prior to the separation of the African populations even though the Africans are known to be extremely diverse. A distinct gradient of decreasing genetic similarity (representing a cline) of Indian populations was observed with the West- and Central-Asian gene
pools as we move eastward or southward from the north-western corridor along the PC2. Figure 3.3 also reveals that the OG population lies on a direct line between North and North-West Indian populations and the Africans, revealing varying levels of admixture between these two broad groups. The distance of the HapMap CEU along the PC2 reduces the possibility that the OGs derive its ancestry from Portuguese traders. European (Portuguese) ancestry was estimated among the OGs, using STRUCTURE and ADMIXTURE and the estimate using both the methods was about 0.03.

To further narrow down on the possible ancestors of OG in India and Africa, PCAs of OG with African populations as well as Dravidian and Indo-European speaking Indian populations that are close to OG was carried out (Figure 3.4), i.e. leaving out the Pakistani populations and the CEU. As observed earlier, PC1 separates the Africans and the Indians (7.95%) and PC2 separates the various African groups. The two pygmy populations (Biaka, Mbuti) and the San of South Africa are well separated from the other African groups, whereas a greater genetic affinity appears to exist between the Mandenka of West Africa, the Yoruba of Central West Africa, and the Bantu speakers from Kenya in East Africa.

Figure 3.4: Genetic relatedness of OG with populations in India and Africa. Principal Component Analysis of 17 Indian (IE speakers and DR speakers) populations and seven African populations from HGDP shows that the OGs are admixed between Indians and West Africans and the Bantu Kenyans.
From **Figure 3.4** it is clear that the OGs lies between the Indians and the Bantu Kenyans, reflecting their varying levels of admixture between these two continental groups. The next step was the identification of the non-African population(s) who might be considered ancestral to the OG. Based on the observations made from **Figures 3.3 and 3.4** as well as distance matrix of EIGENSTRAT and Reynold’s distances between the populations, search of ancestral populations was reduced to the following: IE-N-LP10, IE-N-LP18, IE-W-LP2, and IE-W-LP4. Though the selected populations might not be the exact populations that contributed to OG’s ancestry, genetic distances between population(s) suggests that the actual ancestors should be genetically closely related to the selected populations.

The PCA with the selected four Indo-European speaking Indian populations along with OG and the two African populations show the separation of the Indo-Europeans and Africans along the PC1 (8.67%) (**Figure 3.5**). The variation between the Indo-European populations is more than that between the two African populations, resulting in a clearer separation of the Indo-Europeans along PC2 (1.44%). The variation along PC2 separates seven individuals of the IE-N-LP18 from the rest of the samples. It was observed that these individuals are distinct from the rest of the group in all subsequent analysis.
3.3.2. Genetic differentiation between the ancestral populations

F_{ST} estimates for the OG population and various other populations contributing to the admixture is shown in Table 3.1. F_{ST} estimates between the African and the IE speaking Indian populations are extremely high (> 0.1), whereas the OG lies in between (<0.05). The large pairwise F_{ST} values between the Indian and the African populations are indicative of the large genetic separation between the populations. Analysis of Molecular Variance or AMOVA [33, 34] was performed with 3 groups. First group had four Indian populations, second group with two African populations while the OG was the third group. The AMOVA results confirmed the observations in Table 1; the percentage of variation among groups was 7.92 with a permutation p-value of 0.01 (Table 3.2).

Table 3.1: Extent of genetic differentiation estimated by Pairwise F_{ST} (X 1000) between the populations.

<table>
<thead>
<tr>
<th></th>
<th>IE-N-LP10</th>
<th>IE-N-LP18</th>
<th>IE-W-LP2</th>
<th>IE-W-LP4</th>
<th>OG</th>
<th>Bantu</th>
</tr>
</thead>
<tbody>
<tr>
<td>IE-N-LP10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE-N-LP18</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE-W-LP2</td>
<td>10</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE-W-LP4</td>
<td>8</td>
<td>11</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OG</td>
<td>43</td>
<td>47</td>
<td>49</td>
<td>42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bantu</td>
<td>115</td>
<td>119</td>
<td>121</td>
<td>113</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>Yoruba</td>
<td>128</td>
<td>132</td>
<td>135</td>
<td>126</td>
<td>41</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 3.2: Extent of genetic differentiation estimated by AMOVA.

<table>
<thead>
<tr>
<th>Sources of variation</th>
<th>d.f. (degrees of freedom)</th>
<th>Sum of squares</th>
<th>Variance components</th>
<th>Percentage of variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among groups</td>
<td>2</td>
<td>40816.413</td>
<td>226.08412 Va</td>
<td>7.92</td>
</tr>
<tr>
<td>Among populations within groups</td>
<td>4</td>
<td>14701.267</td>
<td>28.20762 Vb</td>
<td>0.99</td>
</tr>
<tr>
<td>Among individuals within populations</td>
<td>131</td>
<td>335986.892</td>
<td>-35.59279 Vc</td>
<td>-1.25</td>
</tr>
<tr>
<td>Within individuals</td>
<td>138</td>
<td>363764</td>
<td>2635.97101 Vd</td>
<td>92.34</td>
</tr>
<tr>
<td>Total</td>
<td>275</td>
<td>755268.572</td>
<td>2854.66996</td>
<td></td>
</tr>
</tbody>
</table>
### 3.3.3. Estimation of individual ancestry proportions

To estimate individual ancestry (IA) of the OGs considering them admixed between Africans (Yoruba and Bantu Kenya) and Indians (IE-N-LP10, IE-N-LP18, IE-W-LP2, IE-W-LP4), we carried out STRUCTURE [29-31] and ADMIXTURE [32] analysis. The amount of African ancestry as estimated by STRUCTURE is 58.7% ± 8.4%, and there was a range of 40% in the OG individuals (Supplementary Figure 3.A). STRUCTURE models LD arising out of admixture between two genetically distinct populations with different allele frequency spectrum, but does not model the existing LD in the ancestral populations. Genotypes that are in strong LD confounds ancestry estimations computed using STRUCTURE. The SNP density of our dataset is unlikely to affect the STRUCTURE results. Also the ancestral populations in our study are very old, which do not have much background LD. However, since results are highly contingent upon the STRUCTURE findings, validation was done with reduced set of ancestry informative markers. A set of 3396 markers having $F_{ST}$ values ≥0.1 among ancestral Africans and Indians were chosen. STRUCTURE analysis was performed using this set of SNPs minimizes the chance of background LD. Estimates of ancestry computed using set of informative markers and complete set of 18,534 markers are highly concordant (Table 3.3, Table 3.4 and Figure 3.6). It is however important to note, when number of clusters in the STRUCTURE run were increased to more than 2, neither the African populations nor the Indian populations separate. Rather seven individuals of the population IE-N-LP18 separate as a population and the likelihood of the data also reduces.

<table>
<thead>
<tr>
<th>Given pop</th>
<th>Inferred cluster 1</th>
<th>Inferred cluster 2</th>
<th>Number of individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>IE-N-LP10</td>
<td>0.009</td>
<td>0.991</td>
<td>24</td>
</tr>
<tr>
<td>IE-N-LP18</td>
<td>0.002</td>
<td>0.998</td>
<td>23</td>
</tr>
<tr>
<td>IE-W-LP2</td>
<td>0.003</td>
<td>0.997</td>
<td>20</td>
</tr>
<tr>
<td>IE-W-LP4</td>
<td>0.001</td>
<td>0.999</td>
<td>19</td>
</tr>
<tr>
<td>OG-W-IP</td>
<td>0.587</td>
<td>0.413</td>
<td>19</td>
</tr>
<tr>
<td>BANTUKENYA</td>
<td>0.96</td>
<td>0.04</td>
<td>11</td>
</tr>
<tr>
<td>YORUBA</td>
<td>1</td>
<td>0</td>
<td>21</td>
</tr>
</tbody>
</table>
Table 3.4: Proportion of membership of each pre-defined population in each of the two clusters using 3396 markers.

<table>
<thead>
<tr>
<th>Given pop</th>
<th>Inferred cluster 1</th>
<th>Inferred cluster 2</th>
<th>Number of individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>IE-N-LP10</td>
<td>0.077</td>
<td>0.923</td>
<td>24</td>
</tr>
<tr>
<td>IE-N-LP18</td>
<td>0.069</td>
<td>0.931</td>
<td>23</td>
</tr>
<tr>
<td>IE-W-LP2</td>
<td>0.068</td>
<td>0.932</td>
<td>20</td>
</tr>
<tr>
<td>IE-W-LP4</td>
<td>0.068</td>
<td>0.932</td>
<td>19</td>
</tr>
<tr>
<td>OG-W-IP</td>
<td>0.585</td>
<td>0.415</td>
<td>19</td>
</tr>
<tr>
<td>BANTUKENYA</td>
<td>0.937</td>
<td>0.063</td>
<td>11</td>
</tr>
<tr>
<td>YORUBA</td>
<td>0.972</td>
<td>0.028</td>
<td>21</td>
</tr>
</tbody>
</table>

Figure 3.6: Summary plot of individual admixture proportions in OG from Indo-European and African ancestral populations. Each individual is represented by a vertical line broken into two coloured segments. Red lines indicate Indo-European and Green indicate African ancestral proportions. The relative proportion of each ancestor in OG and also in the ancestral populations is represented with length proportional to each of the inferred clusters. (a) Analysis of admixture using the program STRUCTURE assuming two ancestral populations and using data on all 18,534 SNPs. (b) Analysis of admixture using the program STRUCTURE assuming two ancestral populations and using data on 3396 AIMs ($F_{ST} > 0.1$).
3.3.4. Linkage Disequilibrium in OG and ancestral populations

It is known that, admixed populations have elevated levels of LD irrespective of existence of LD in ancestral populations. The LD in the OG population is observed to be much higher than the Indian and/or the ancestral African population (Figure 3.7).

The pairwise LD analysis for the ancestral populations ($r^2$ threshold > 0.2) resulted in 360 marker pairs in the Indian population and 329 marker pairs in the African population. In contrast, in OG, there were 1100 marker pairs with $r^2$ > 0.2. The average $r^2$ drop below 0.05 for Africans and Indians within 250 Kbp, whereas the average LD is steady and above 0.1 for the OG even at distances larger than 800 Kbp. The extremely long-ranged LD among OG, compared to ancestral populations, indicates that the admixture event is relatively recent.

3.3.5. Gene ontology analysis of ancestry informative markers

It would be interesting to see if there were some biological processes that were selectively enriched in the admixed populations from either of the ancestors. Considering the SNPs which have an $F_{ST}$ value > 0.1 between the two ancestral populations 3396 of the 18,534 SNPs were selected for functional analysis. 1218 SNPs
were filtered out from this analysis as their frequencies in the OG population were within 5% of the expected frequency, which is the ancestry proportionate weighted average of the allele frequencies of the two ancestral populations. The remaining SNPs were classified into two groups of 1240 and 938 SNPs based on their closeness, in terms of allele frequency, to the Indian and African ancestral populations respectively. Analysis of gene classes in these groups revealed significant enrichment of cadherins, potassium channels, membrane proteins and solute carriers as well as protein kinases from the group close to IE and kinases and immune related genes from the group close to African ancestry. Further, Functional Annotation Clustering (FAC) revealed significant enrichment of processes related to axonogenesis and potassium transport. SNP frequencies of the genes involved in these processes are close to the Indian ancestral population (Table 3.5). However, FAC did not reveal any specific enrichment of the processes contributed by the other group.

Table 3.5: Functional annotation of genes encompassing the ancestry informative markers which are closer in allele frequency to the Indian ancestral population.

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>P value</th>
<th>Fold Enrichment</th>
<th>Bonferroni</th>
<th>Benjamini</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Enrichment Score: 4.324</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0007409−axonogenesis</td>
<td>16</td>
<td>9.43E-06</td>
<td>4.063</td>
<td>0.015</td>
<td>0.015</td>
<td>0.016</td>
</tr>
<tr>
<td>GO:0048667−cell morphogenesis involved in neuron differentiation</td>
<td>16</td>
<td>2.44E-05</td>
<td>3.752</td>
<td>0.039</td>
<td>0.008</td>
<td>0.041</td>
</tr>
<tr>
<td>GO:0048812−neuron projection morphogenesis</td>
<td>16</td>
<td>3.04E-05</td>
<td>3.682</td>
<td>0.048</td>
<td>0.008</td>
<td>0.051</td>
</tr>
<tr>
<td>GO:0000904−cell morphogenesis involved in differentiation</td>
<td>16</td>
<td>1.42E-04</td>
<td>3.214</td>
<td>0.207</td>
<td>0.023</td>
<td>0.238</td>
</tr>
<tr>
<td>GO:0031175−neuron projection development</td>
<td>16</td>
<td>2.41E-04</td>
<td>3.063</td>
<td>0.325</td>
<td>0.030</td>
<td>0.403</td>
</tr>
<tr>
<td><strong>Enrichment Score: 3.240</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potassium transport</td>
<td>10</td>
<td>2.95E-04</td>
<td>4.672</td>
<td>0.097</td>
<td>0.011</td>
<td>0.403</td>
</tr>
<tr>
<td>Potassium</td>
<td>10</td>
<td>5.79E-04</td>
<td>4.264</td>
<td>0.182</td>
<td>0.012</td>
<td>0.788</td>
</tr>
<tr>
<td>GO:0030955−potassium ion binding</td>
<td>10</td>
<td>0.001112</td>
<td>3.871</td>
<td>0.437</td>
<td>0.062</td>
<td>1.596</td>
</tr>
</tbody>
</table>
3.4. Discussion

In this study I have dissected the ancestry and genetic structure of an Afro-Indian population residing in Gujarat using a set of genome-wide 18534 markers. Though it is well acknowledged that the OG population has an African origin, their specific ancestry and time of spread in mainland India has been enigmatic. I tried to elucidate the genetic structure of this population using a set of populations from the CEPH-HGDP panel, HapMap and 26 diverse Indian populations. In an earlier study on 55 Indian populations using a set of 405 SNPs it was observed that the OG population was distinct from the rest of India populations [2]. Extending the analysis to genome-wide markers in a subset of these populations further substantiated earlier observations (Figure 3.2). This is also supported by the recent observation of Reich et al [5].

A prior analysis of genetic structure among the African populations included in the HGDP based on 377 autosomal STR loci was able to define distinct genetic clusters for the Biaka, Mbuti, and San; however, the study lacked the power to differentiate the Mandenka, Yoruba, and Bantu groups [40]. In contrast, greater resolution of African ethnic groups, particularly for the Mandenka and Yoruba, was possible in multiple recent studies [11, 41, 42]. This study suggests that the African slaves brought in India are unlikely to be of diverse origin since PCA revealed that they are closer to Bantu Kenya and YRI. This is unlike the multiple migration and massive slave trade that happened in the new world. Hence, though it is of interest to compare African admixture estimates to descriptions of proportional representation of various African groups to the Middle Passage and slave trade occurring in post-Columbian America; it is unlikely that the situation will be replicated in the Indian context. However, in absence of data from population groups representative of South-Eastern and other parts of Southern Africa, their genetic representation in OG still remains a possibility.

It is important to note that considerable migration has occurred among African ethnic groups over the past three millennia or more. For example, the two Bantu groups included in our analysis originated from a more-central African location (Nigeria-Cameroon) several millennia ago, making precise geographic localization of African ancestry difficult [42, 43]. This difficulty is also reflected in the close genetic relationships amongst the various West, West Central, and South West African groups,
who also show considerable overlap in terms of mtDNA haplotypes vis-à-vis the autosomal genome [41]. Recent large scale studies on the African genetic diversity also substantiate the closeness of the Bantu and Yorubans and have only limited representation of South and Southeastern population groups [41]. Previous genetic study on the OG population had suffered from similar limitations and has rarely tried to address their ancestry [23]. This is because they either lacked data on reference population or had very little genetic data, often from a single marker and a few loci. Results from this study are based on examination of the entire autosomal genome and, therefore, provide a more-robust picture of the admixed African ancestry of individual African Indians compared to prior analyses, which focused on only a single locus (mtDNA or Y chromosome).

This exploratory analysis looking for the non-African component of the ancestry of OG started in an agnostic fashion. Six North Indian IE speaking populations, five IE speaking populations from West India along with three from the East were included. Also, two Dravidian populations and one IE population from North-East were included. From the most comprehensive study of human genetic diversity of Indian population (at least in terms of number of population represented) [2], only Tibeto-Burman (TB) speaking populations of India were excluded. The TB who primarily inhabit the North-Eastern parts of India are genetically close to East Asians and comprise a distinct genetic pool [1, 2]. They do not overlap the domain of the OG either geographically or linguistically or otherwise. The present analysis corroborates with observations by Reich et al. [5] and other groups [1-3], a distinct gradient of decreasing genetic similarity (representing a cline) of Indian populations with the West- and Central-Asian gene pools as we move eastward or southward from the north-western corridor (Figure 3.4). It also shows that if CEU, along with Pakistani and IE and DR populations from all over India are included, the genetic variation is more than that observed in the HGDP African populations. However, when I narrow down agnostic search to include only the Indian populations, the intra-African variation is much larger. The oral history states that African ancestors to the OG population were brought into India by the Portuguese traders. It is possible that a very small fraction of the non-African ancestry can actually be derived from the Portuguese traders. Using both ADMIXTURE and STRUCTURE,
CEU ancestry was estimated to be around 3%. Given the marker density in our dataset it is difficult to attribute whether that small proportion is ‘real’ or an artifact.

Prior studies on African populations suggest close genetic kinship among various West, Central West, and South-West African ethnic groups [41]. It is to be noted here that identification of ‘exact’ ancestors to admixed population(s) is a problem which is impossible to address. This applies specially for a population like the OG which has remained small and has undergone random genetic drift and possible selection. I do not claim here that the four Indian and two African populations which were introduced as possible ancestors to OG are their ‘exact’ ancestors. Rather, it can be claimed that the differences between the two ancestral populations contributing in OG admixture are genetically so diverse that our choice is a good approximation for all practical purposes. The ancestry estimates are also largely independent of the number of markers used and irrespective of the choice of markers (i.e., entire genome or ancestry informative markers), estimates are pretty robust to the choice.

The $F_{ST}$ estimates between the African and the IE speaking Indian populations are extremely high (> 0.1). The large pairwise $F_{ST}$ values between the Indian and the African populations are indicative of the large genetic separation between the populations. It is also indicative of the fact that there is expected to be a large number of ancestry informative markers, which ensures a relatively easy and efficient study design for MALD. It was observed that 3396 (>18%) of 18,534 SNPs to have $F_{ST}$ values larger than 0.1 between the Indian and the African ancestral populations indicative of the large genetic separation between the populations. This as well as the distribution of the AIMS on genes association in different diseases as listed in GAD (Supplementary Figure 3.B) also indicates that the OG population is likely to be extremely informative in MALD.

The extent of LD is contingent upon the allele frequency difference between the ancestral populations as well as the number of generations elapsed after the admixture event has taken place and it rapidly decrease with each passing generation [12, 13, 44]. The long range LD among the OG is indicative of very recent admixture. Since there is little prior study of the African diaspora in the ocean of Indian population diversity it is difficult to state when the African ancestors to OG started settling here. However, it is
likely that there were limited mate exchange with native populations, till recent times, probably because of sociological constrains. This is also evident from individual estimates of Indian ancestry which shows that the Indian contribution to the admixture is approximately 40%. Some historians argue that in the 19th century, Zanzibar emerged as the hub for the distribution of African slaves to Arabia, southern Persia and probably western India. Even after the nominal abolition of the slave trade by the British, a small number of male and female African slaves continued to be shipped to the western coasts of South Asia, especially to Makran and Gujarat, where they were mostly employed as servants and bodyguards at the courts of local rulers. The long-range LD is also possible if the 19th century Zanzibar residents are the African ancestors to the OG. The genetic similarity of the African ancestors of the OG to the current West Africans can still happen due to the wide spread of Bantu speakers throughout Africa and their genetic homogeneity [42].

Migrations of Africans into mainland India, brought individuals from different continents into close physical proximity and inter-mating between migrant and indigenous populations. This meant sudden confluence of geographically diverged genomes with novel environmental challenges. These unprecedented events brought together genomes that had evolved independently and optimized to different continents and conditions for tens of thousands of years and presented new environmental challenges for the indigenous and migrant populations, as well as their offspring. These circumstances provided novel opportunities for natural selection to occur and perhaps resulted in large deviations at specific locations from the genome-wide ancestry distribution [9, 10]. As it has already been shown that the OG is a relatively recent admixed population, large deviations at specific locations from genome-wide ancestry distribution was not expected. However, it was worthwhile to examine that whether there are enriched biological processes in OG which were retained from either of the ancestors. Search for functional enrichments was directed at the AIMs that were associated with genes and whose frequency in OG was close to either of the ancestral populations. Significant enrichment of processes related to ion channel activity and cadherin genes was observed, which in terms of its genotypic spectrum is close to the IE ancestors (Figure 3.8).
Figure 3.8: Genotype distribution of AIMS related to ion channel activity and cadherin genes in OG and its ancestral populations.

The genotypes of various AIMS depicted in columns are represented for each individual in a row. Heterozygous genotypes are represented in pink and the two homozygous genotypes are represented in green and cream for each of the markers. The genotypes of 19 individuals of OG and representative Indian (IE-N-LP4) and African (YRI) ancestral populations are depicted. The genotypes of OG are markedly similar to the Indian IE ancestor than the African ancestor.

Selection in ion-channel genes among populations of African ancestry has been a long-term global enigma [45]. However, the fact that the population resides in an extreme saline region of the country and has shown deviations in these genes was intriguing and compelling to speculate that this finding is biologically relevant. This is especially interesting in the light of a recent GWAS study of hypertension and blood pressure in African Americans where similar family of genes related to ion channels, cadherins and calmodulins have been implicated [46].

Ramana et al. studied the variation in the Y chromosome among the OG [23] and found that there is considerable infusion of Y chromosomes from different Indian caste populations into the gene pool of OG. Although the African Indian population was sampled from a different geographical location, it probably shares a common history.
with the population sampled in this study. Despite the Y-chromosomal variation, there is little chance that the maternal founding gene pool for the OG population is large. The population also lives in isolated small endogamous groups, which is the likely cause of the deep-rooted founder effect. The population history hence resembles, in terms of forces shaping genetic architecture, the European Roma population and the Ashkenazi Jews, who have long served as a model population for identification and mapping of founder mutations and diseases [47, 48].

The overarching goal of identifying an admixed population lies in potential of the population in mapping disease causing mutations. Admixture mapping is based on the hypothesis that differences in disease rates between populations are due in part to frequency differences in disease-causing genetic variants. In admixed populations, these genetic variants occur more often on chromosome segments inherited from the ancestral population with the higher disease variant frequency [13]. Hence, the chances of successfully mapping disease causing variants are vastly improved if the divergence between the ancestral populations is large. It also takes advantage of long-range haplotypes that are generated by gene flow among recently admixed ethnic groups. The chance further increases if there is a large difference in the prevalence of the disease between the ancestral populations. The extent of LD also ensures that the admixture event is recent and large segment of the ancestral sequences are inherited from the ancestral population to the admixed population. It can be speculated that large divergence of the two ancestral populations and the recent admixture makes OG a candidate population for admixture mapping. Around same time such findings were also elaborated by other group [49].
3.5. References


