Chapter 6: Population relationships as inferred from classical HLA genes

In this chapter, our aim is to investigate the extent to which allele and haplotype frequency distributions at the classical HLA loci can be used to infer genetic relationships between the population datasets collected as part of the 13th International Histocompatibility Workshop (13th WS) (1). Where possible, these relationships will be used to make inferences about the demography and history of these populations, and to test specific hypotheses about human history and migration and colonization patterns.

While there are several graphical means of presenting inferred population relationships (e.g., genetic distance tables, phylogenetic dendrograms, principal component plots), dendrograms (a.k.a., trees) are presented here. Trees can be generated using population-level frequency data (population trees) or individual-level sequence data (gene sequence trees) for a particular gene. Only population trees are presented here. While the topology of an inferred tree is influenced by the molecular changes that occurred in the history of a particular gene, genetic drift and selection, as well as population demography, migration, fusion and admixture, can influence the topology as well. With the assumption of selective neutrality and particular mutation and fixation rates, gene sequence trees can be interpreted under a molecular clock model and can be used to make inferences about the time since sequences diverged from each other; in some instances, the topology of these trees can be used to reconstruct the nucleotide sequences of ancestral alleles. Whereas sequence trees can sometimes be viewed as historical gene genealogies, population trees should be considered as graphs of the general trends in relationships between modern populations, which can change in ways (e.g., admixture, splitting, fusion, bottlenecks, etc.) that nucleotide sequences cannot. Unlike the case with sequence trees, the branch lengths in population trees cannot be directly interpreted in terms of time, as demographic effects (e.g., gene flow and population size) must be...
considered as well. In particular, it should be noted that the relationships represented by population trees are generally representative of the first (and sometimes second) principal components of the allele/haplotype frequency data.

Sequence trees are often rooted by using an outgroup (i.e., a non-human primate sequence in the case of human sequences) or by mid-point rooting (where the root is placed halfway between the two most distant taxa). All population trees are presented here as unrooted, as it is difficult to know exactly how or where to place the root. For example, the influence of the number of alleles in a population (k) on branch lengths raises questions about the effectiveness and appropriateness of mid-point rooting. Similarly, because these trees are based on frequency distributions, the lack of a non-human population sharing allelic and haplotypic diversity with human populations makes outgroup rooting difficult. While some information is absent from unrooted trees (i.e., the temporal ordering of population relationships), these trees remain useful representations of the relationships between populations, permitting the inference of relationships based on shared history or gene flow.

It should be noted that the allele and haplotype frequency based tree building algorithms used here do not take into account the sequence relationships among alleles, as in the case of sequence tree analyses. For alleles that diverged before the migration of human populations out of Africa (ca 100,000 years ago), this distinction should not be problematic (i.e., spurious genetic and inferred historical relationships should be minimal). However, in some cases, alleles and haplotypes that have been generated relatively recently and are present only in one population at high frequencies may create such problems (e.g., long branch lengths resulting from infinite genetic distances between populations). Similar difficulties result from isolated populations with low values of k, in which a few alleles are subject to genetic drift. These issues are discussed below.

In addition, analyses of sequence trees and population trees constructed using data from the same genes, or comparisons of trees made for the same populations using data from different genes can result in different inferences. For example, sequence trees constructed using mitochondrial DNA (mtDNA) and Y-chromosome markers for Polynesian, western Austronesian, and Southeast Asian populations (2–4) support the “entangled-bank” model for the colonization of the east Pacific, in which the ancestors of Polynesian populations migrated slowly eastward from Southeast Asia, acquiring Austronesian mtDNA and Y-chromosome markers as the result of admixture with those populations. Population trees constructed using DRB1 and DRB1-DQB1 haplotype frequency distributions (5) support the “fast-train” model of Polynesian colonization, in which Polynesian ancestors moved rapidly from Southeast Asia to the east Pacific, with little admixture with intervening populations to bring about changes in allele and haplotype frequency distributions. Such distinctions likely stem from differences in the modes of evolution of the genes being studied; the mtDNA and Y-chromosome markers studied are not under selection and are free to accumulate distinguishing nucleotide substitutions relatively rapidly in comparison to HLA alleles, most of which predate the “out of Africa” expansion (as opposed to having been recently generated) and are shared between Polynesians, Southeast Asians, and Austronesians. Where the transmission of population-specific mtDNA and Y-chromosome sequences from one population to another will be apparent in a phylogenetic study of individual sequences, the admixture of individuals with similar alleles between two otherwise unrelated populations may not have a dramatic effect on the overall allele frequency distributions.

Finally, the classical HLA loci (with the exception of DPB1) are believed in general to be under the influence of balancing selection, which will offset the effects of genetic drift (e.g., by maintaining novel and low frequency alleles in large populations) and of directional selection (e.g., by preventing the predominance of a single allele) on allele frequency distributions. It seems likely that the maintenance of similar allele frequency distributions between related populations by balancing selection contributes to the utility of the HLA loci for making phylogenetic inferences. However, some populations (especially small populations) may still experience extreme examples of genetic drift (e.g., population bottlenecks) and directional selection (e.g., the predominance of an allele observed in many populations in response to a local, active pathogen, or the high frequency of a novel allele in response to selection for sequence diversity), with large changes to allele frequency distributions as a result. In these cases, dramatic departures from the allele frequency distributions of ancestral and related populations will likely result in spurious population relationships being inferred from population trees.

Methods

Each population dataset generated as part of the 13th WSS was assigned to one of 11 categories (SSA, NAF, EUR, SWA, SEA, OCE, AUS, NEA, NAM, SAM, and OTH) describing its region of origin, as described elsewhere in this volume (1). With the exception of populations in region OTH (Other), which were excluded from analysis, these regional designations were ignored for the calculation of genetic distances and the gener-
The programs in the PHYLIP (v 3.62b) software suite were used to calculate genetic distance matrices and construct trees from these distances (6). Nei’s Standard Genetic Distance (7) was calculated in GENDIST for all populations with 2n > 49 and k > 6 at each locus (see below), with the exception of the DQB1 locus. Because the number of alleles at the DQB1 locus is considerably lower than at the other loci, populations were not excluded from analysis on the basis of the number of alleles (k) at this locus. Values for infinite genetic distance (resulting from comparisons between populations with low numbers of alleles or haplotypes in common) were not modified from those assigned by the GENDIST program (−1), although the percent of infinite distances in each genetic distance matrix was calculated for each tree.

Trees were constructed using the Neighbor Joining (NJ) algorithm (8) in the NEIGHBOR program, with randomized input orders (Jumble option) for branch addition. Trees were unrooted in every case. Using these methods, a tree can be generated using allele frequency data for a single locus only (single-locus trees), using data from multiple loci (multi-locus trees) with the allele frequencies at each locus considered independently, and using frequency data for individual haplotypes (haplotype trees). No trees were generated using data for multiple haplotypes. Single-locus trees were generated for the HLA-A, C, B, DRB1, and DPB1 loci, and multi-locus trees were generated for the following combinations of loci: HLA-A & B; HLA-A, B & DRB1; HLA-C & B; and HLA-DRB1 & DQB1. Haplotype trees were generated for the following haplotypes; HLA-A:B; HLA-A:B:DRB1; HLA-C:B, and HLA-DRB1:DQB1. Not all trees generated were shown or discussed for the reasons described below.

For the purpose of evaluating results, a uniform set of criteria was used to accept or reject a given tree for further discussion. Studies of population relationships at non-MHC loci have generally shown a close correlation between genetics and geography, so that populations tend to share similar allele frequencies with their neighbors on a local level (9, 10). The reasons for this observation are arguable, but likely include varying degrees of shared ancestry, recent admixture, local selection, and isolation by distance. Similar correlations between HLA genetics and language are presented in this volume (11). Therefore, we expect that most populations will be more closely related to their neighbors (populations in the same global region) than to non-neighboring populations, and that most relationships will corroborate geographic, historical, anthropologic and linguistic evidence as a result. If more than 6% (rounded up) of the intraregional population relationships in a tree do not meet this expectation (e.g., an Australian population is placed in a group with European populations, with no other Australian populations near by), that tree is rejected as invalid. For a given marker, some populations may have relatively similar allele frequencies (and thus a small genetic distance) due to chance rather than common ancestry, gene flow, or other historical relationships. If the populations within a given region are split into two distinct groups, the relationships between the populations within each group are considered to meet the expectation of geographic, historical, anthropological and linguistic correlation (e.g., the populations in region SWA, below). For the purpose of these evaluations, North and South American populations were considered to be in the same region, even though they are assigned to and are drawn as belonging to separate regions. It follows that these criteria only apply to those global regions represented by more than one population in a given tree, and that they are necessarily conservative; if no expectations are met, it is difficult to say which relationships are genuine, and which might be spurious. Finally, population relationships, however unexpected, that are repeatedly observed in trees derived from different markers, clearly merit serious consideration as reflecting historical patterns rather than as a spurious artifact of the tree-building process.

All populations were included in preliminary analyses (regardless of their size and number of alleles), and all trees but the DRB1:DQB1 haplotype tree were rejected using the criteria described above. Upon careful examination of the dataset, we observed that most of the presumably spurious population relationships involved populations with low values of 2n and k. Consequently, we have restricted the analyses presented here to populations with values of 2n > 49 at all loci and k > 6 at each loci except DQB1. It seems likely that more stringent criteria (e.g., including only those populations with 2n > 75 and k > 8 at each locus) would result in the generation of a greater number of acceptable trees, but this would dramatically reduce the number of population datasets available for analysis.

Results

Applying the 6% criteria described above, the single-locus HLA-B tree, the multi-locus HLA-C & B tree, and the HLA-DRB1:DQB1, HLA-A:B, and HLA-C:B haplotype trees were ac-
accepted as being valid for discussion. These trees are shown in Figures 1–5. While not all of the population relationships in these figures correspond to the designated global regions, several of the groups containing such non-corresponding populations are consistent with geography and specific anthropological hypotheses. In addition, some of these unexpected relationships appear in multiple trees and are therefore supported by the data from multiple genes and haplotypes.

In Figure 1 (HLA-B alleles), the placement of the “North American Amerindian” population (in a group containing populations from Sub-Saharan Africa (SSA), North Africa (NAF), Europe (EUR), and Southwest Asia (SWA)) appears to be spurious, as this relationship is not consistently supported in other trees. This “North American Amerindian” dataset is an admixed population sample that represents individuals from many native North American populations, and the allele frequency distributions in this dataset may not reflect those of any actual population. Other Native American populations in this tree cluster together and are closest to the Tuva, Okinawans and Koreans. In almost every tree (see below), the Ugandan population is the closest of the SSA populations to populations from NAF, EUR and SWA, so this relationship is not inconsistent with the other analyses. The Omani population (from the Arabian peninsula) is close to populations from EUR in this tree, and appears in a spurious group in Figure 4 (see below). This population was placed close to other European populations in trees that were rejected as invalid, so the significance of the placement of this population is unclear. The placement of the Georgian population, closer to populations in SWA than EUR in this and in other trees (see below) may suggest that the border between EUR and SWA needs to be modified. Similarly, the close proximity of the Malay population to populations from Oceania (OCE) suggests that the border between Southeast Asia (SEA) and OCE should be reconsidered. The placement of the PNG Highland population in a group with the populations of Australia (AUS) is predicted by the “Sahul-hypothesis” of common ancestry between Papua New Guinea Highland populations and aboriginal Australian populations (see below).

In Figure 2 (HLA-C & HLA-B alleles), the placement of the Canoncito and Bari outside of North and South America (NAM and SAM) is likely spurious. The close proximity of the Finnish population to populations in Northeast Asia (NEA) may represent historical gene flow from that region, as a similar relationship is seen in Figures 1 and 5. As in Figure 1, the placement of the Georgian population in a group with the Kurdish population, and the Malay population in a group containing populations from OCE suggests a reevaluation of the border between EUR and SWA, and between SEA and OCE.

In Figure 3 (HLA-DRB1:DQB1 haplotypes), as in Figure 1, the placement of the PNG Highland population in a group with populations from AUS again supports the “Sahul-hypothesis” that PNG Highland and Aboriginal Australian populations are descendants of the original colonists of Sahul, the land-mass that consisted of Australia, Tasmania and PNG during the Last Glacial Maximum (LGM). Following the separa-
Figure 2. Tree constructed using HLA-C and HLA-B allele frequencies in 53 population samples. This figure depicts a Neighbor Joining tree constructed using the HLA-C and -B allele frequency distributions of 53 13th WS population samples with sample sizes (2n)=49 and numbers of alleles (k)≥6 at each locus. Populations in the same global region are contained within polygons of the same color, and are indicated in text of the same color. Correspondences between polygon and text colors and global regions (Sub-Saharan Africa, North Africa, Europe, Southwest Asia, Southeast Asia, Oceania, Australia, Northeast Asia, North America, and South America) are shown in the inset box. The inset scale describes genetic distances of 0.1.

In Figure 4 (HLA-A:B haplotypes), the placement of the Mandenka, Yuendumu and Omani populations is likely spurious. As in Figures 1 and 2, the close proximity of the Malay population to the Indonesian and Filipino populations suggests that the boundaries of global regions may be different for class I versus class II loci. In some of the class I trees discussed here, the Malay population seems distinct from other Southeast Asian populations, and closer to Oceanian populations (e.g., Indonesians).
Figure 3. Tree constructed using HLA-DRB1:DQB1 haplotype frequencies in 28 population samples. This figure depicts a Neighbor Joining tree constructed using the HLA-DRB1:DQB1 haplotype frequency distributions of 28 13th WS population samples with sample sizes \((2n)>49\) and numbers of alleles \((k)>6\) at the DRB1 locus. Populations in the same global region are contained within polygons of the same color, and are indicated in text of the same color. Correspondences between polygon and text colors and global regions (Sub-Saharan Africa, North Africa, Europe, Southwest Asia, Southeast Asia, Oceania, Australia, Northeast Asia, North America, and South America) are shown in the inset box. The inset scale describes genetic distances of 0.1.

This suggests a re-evaluation of the boundary between SEA and OCE. The placement of the American Samoa population in a group containing populations from SEA is predicted by the "Fast-train model" hypothesis for the colonization of Polynesia.

In Figure 5 (HLA-B:C haplotypes) the position of the American Samoa population again supports the "Fast-train model" hypothesis for the colonization of Polynesia. As was the case in Figures 1 and 2, the inclusion of the Malay and Indonesian populations in a group suggests that the boundary for SEA and OCE (which runs between the Malay peninsula and Indonesia) should be re-evaluated.

Of the remaining trees that were generated, several (HLA-A, HLA-DRB1, HLA-A & B, HLA-A, B, & DRB1, and HLA-DRB1 & DQB1) came close to being considered valid, with between 6% and 10% of relationships failing to meet expectations of geography, history, anthropology and linguistics (not shown). The correspondence to these expectations was lowest in the HLA-A:B:DRB1 haplotype tree (not shown) with 56% of the population relationships appearing to be spurious. This observation reflects the lower linkage disequilibrium characterizing these 3 locus haplotypes that span around 2 Mb. Recombination creates new A:B:DRB1 haplotypes, generating great haplotypic diversity, and many of these haplotypes will be population or region specific. The percent of infinite dis-
Figure 4. Tree constructed using HLA-A:B haplotype frequencies in 57 population samples. This figure depicts a Neighbor Joining tree constructed using the HLA-A:B haplotype frequency distributions of 57 13th WS population samples with sample sizes (2n)$>49$ and numbers of alleles (k)$>6$ at each locus. Populations in the same global region are contained within polygons of the same color, and are indicated in text of the same color. Correspondences between polygon and text colors and global regions (Sub-Saharan Africa, North Africa, Europe, Southwest Asia, Southeast Asia, Oceania, Australia, North Asia, North America, and South America) are shown in the inset box. The inset scale describes genetic distances of 0.1.
Figure 5. Tree constructed using HLA-B:C haplotype frequencies in 53 population samples. This figure depicts a Neighbor Joining tree constructed using the HLA-B:C haplotype frequency distributions of 53 13th WS population samples with sample sizes ($2n$) $>49$ and numbers of alleles ($k$) $>6$ at each locus. Populations in the same global region are contained within polygons of the same color, and are indicated in text of the same color. Correspondences between polygon and text colors and global regions (Sub-Saharan Africa, North Africa, Europe, Southwest Asia, Southeast Asia, Oceania, Australia, Northeast Asia, North America, and South America) are shown in the inset box. The inset scale describes genetic distances of 0.1.
stances in the distance matrix was highest for this tree (39%), and the invalid topology of this tree is likely due to the extremely high number of A:B:DRB1 haplotypes (2750) that were included in this analysis; while only 25 populations were included in this tree, such a large number of haplotypes were considered for each population that there are many instances in which the haplotypes observed in a pair of populations were almost mutually exclusive, confounding meaningful comparison. The percent of infinite distances in other trees ranged from 0% for all multi-locus trees to 7% for HLA-A:B haplotypes, and it seems unlikely that distance matrices in which large fractions of the distances are infinite will yield useful and meaningful trees.

Several features are common to the trees in Figures 1–5. Populations in SSA are usually (with the exception of Figure 1) observed in a single group, with the Ugandan population closest to non-Sub-Saharan African populations. Depending on the locus, populations in NAF, EUR, or SWA are closest to the Sub-Saharan African populations. In many instances (especially for the DRB1:DQB1 and C:B haplotype trees, but to a much lesser extent for the HLA-B tree), the branch lengths separating the populations in these four regions are short in comparison to the remainder of the world, suggesting both a greater differentiation of allele and haplotype frequency distributions in populations from SEA, OCE, AUS, NEA, NAM, and SAM, and a much greater degree of similarity between populations in Africa and populations in Europe.

The populations in SWA are of particular interest as they are sometimes observed in distinct eastern and western groups. In the HLA-B tree (Figure 1), the Arab Druze, Israeli Jewish, Kurdish, New Delhi, Tamil and South Indian populations are observed in a single group (along with the Georgian population), while in the HLA-B:C haplotype tree (Figure 5) the Arab Druze, Israeli Jewish and Kurdish populations are observed in a group (along with the Georgian population) that branches off from populations in EUR, while the New Delhi and Tamil populations are observed in a group that branches off from SEA. This may be due to westward gene flow of C:B haplotypes from SEA into SWA. Analyses of haplotypes have proven useful for inter-regional comparisons of populations (5, 12), as haplotypes can sometimes serve to “subdivide” alleles, distinguishing variants that would otherwise appear identical. In addition, slightly different relationships between the populations in SWA and those in NAF and SEA are seen in the HLA-A:B haplotype tree (Figure 4), where the Kurdish population is included in the SWA group that branches off of SEA. A high-density sampling of populations from SWA and bordering populations from EUR (e.g., the Georgian population) and SEA will perhaps clarify the relationships within this region, in future studies. The inclusion of the Georgian population in groups containing populations from SWA suggests that the boundary of NAF and SWA should be reconsidered for future studies. It is unfortunate that the Turkish population was only included in the DRB1:DQB1 haplotype tree as this population is found along the western border of SWA. For this reason, future studies should focus on complete class I and class II typing for all populations.

Approximately 14 populations from the island of Taiwan are included in SEA. In almost all trees including Taiwanese populations, the populations of SEA are observed in two distinct groups; a “continental” group containing Han Chinese, Thai, North American Asian, Malay, and Minnan and Hakka (two non-indigenous Taiwanese Han Chinese populations) and a Taiwanese group, containing the aboriginal Taiwanese populations. This is perhaps best exemplified in Figure 5. This pattern is particularly interesting in that the branch lengths of Taiwanese populations are as long in many cases as are those for populations in entire other regions (c.f., SSA in Figure 5), indicating a tremendous diversity of allele and haplotype frequency distributions in an isolated area. It is not clear if this level of haplotype frequency diversity is specific to Taiwan or if similar patterns would be seen whenever allele and haplotype frequencies for a large number of populations sampled from a small area are analyzed in this manner. In the former case, it might be prudent to designate Taiwan as a separate global region for future studies.

Of the populations in NEA, only the Okinawan, Tuva and Korean populations are included in the trees presented here. The populations in this region are quite distinct from the populations in SEA. This is particularly striking in the case of the Okinawan population, as the Okinawan islands lie considerably south of Korea and the Republic of Tuva, and suggests that this population has been isolated for considerable time.

In many instances, groups containing populations from NAM and SAM branch off from populations in NAF, consistent with the theory that the Americas were colonized from northeast Asia. This is the case in Figures 1, 4 and 5, and is largely the case in Figure 2 with the exception of the Finnish population. The two Guarani populations (Kaiowa and Nandeva) are found in a group together (sometimes to the exclusion of other populations from NAF and SEA) in class I analyses (Figure 1 and 4) but not in class II analyses (Figure 3). This could be due to the relatively large number of putatively novel HLA-B alleles observed in South American populations (13). HLA-B*1504,
*3905, *0404, and *3505 are the highest frequency B alleles in the Guarani populations, with frequencies >10% in most cases. It seems likely that the reduced allelic diversity, and the presence of putative novel and population-specific alleles, of populations from NAM and SAM frequently confound genetic distance analysis; many of the pairs of populations with infinite genetic distances include populations from these regions. This reduced diversity is likely also the reason why populations from NAM and SAM cannot easily be assigned to region-specific groups based on HLA frequency data.

Conclusions

Based on these phylogenetic analyses, it seems possible to distinguish “older world” populations in Africa, Europe and the “Middle East” (western Southwest Asia) from the rest of the world. These areas have been inhabited for considerable periods of time, and are always observed in a distinct group. Outside of this region, the topology of the trees depends to some extent on the loci analyzed. Northeast Asians and North and South Americans are generally closest to this “older world” group, followed by Oceanians, eastern Southwest Asians, or Southeast Asians. Australians tend to be quite distinct from the other populations, consistent with the long period of isolation (~50,000 years) these populations have experienced. Higher-resolution and complete (class I and class II) typing of these populations, and an increased density of sampling of populations along the borders of these regions will likely help in clarifying the relationships between these regions in future studies.

In the case of haplotypes in strong linkage disequilibrium (DRB1-DQB1 and C:B haplotypes), branch lengths between populations in more recently colonized regions of the world (the “newer world” of North and South America, Oceania, and Australia) are considerably longer than branch lengths in the rest of the world (Sub-Saharan Africa, North Africa, Europe, Southwest Asia and Northeast Asia) suggesting a greater differentiation of allele and haplotype frequency distributions in more recently colonized regions, possibly resulting from the influence of successive founder effects and genetic drift on the small populations that colonized these regions.

Populations of Southwest Asia often fall into two distinct sub groups, a “West Asian” or “Middle East” region that shares affinity with Europe and North Africa, and a “South Asian” or “Indian Subcontinent” region that shares affinity with South-East Asia. This distinction results from the presence in Indian populations of alleles and haplotypes that are common in South-East Asia, but which are not common in “Middle Eastern” populations, and challenges the concept that South Asian populations are recently derived from western Asia and Europe.

The inclusion of a large number of Taiwanese populations has presented an opportunity to compare allele and haplotype diversity in a small region with the rest of the world. The somewhat surprising result demonstrates the degree to which isolated neighboring populations can present distinct allele and haplotype frequency distributions. Future studies should focus on similar high-density sampling of isolated populations.

While these results are generally similar to phylogenetic analyses of non-MHC loci allele frequencies published by others (10), it is important to note that the population relationships revealed through HLA analysis rely on data for one or two loci, while numerous (50 or more) non-MHC loci must be combined in similar analyses to achieve comparable results. This difference is due to the extremely polymorphic nature of the HLA loci, and it is this polymorphism that makes these loci extremely useful for examining population relationships. Finally, future studies should focus on surveying those regions of the world that have been poorly studied (e.g., Sub-Saharan Africa, and Central, South and Northeast Asia) here, with a particular focus on populations at the borders of these global regions.

References


