

Stem Cell Donor Matching for Patients of Mixed Race

Ted Bergstrom, Rod Garratt, and Damien Sheehan-Connor

December 28, 2009

1 Background

Patients with leukemia and other blood diseases frequently stand a good chance of recovery and a return to normal life if they receive a stem cell transplant from a living donor. In the absence of a transplant, their survival prospects are grim. For a transplant to be successful, the human leukocyte antigens (HLA) of the donor and recipient must be a close genetic match. This requires that donor and recipient have the same alleles at each of three genetic loci.¹ Each potential donor or recipient has two copies of the gene at each locus, one inherited from each parent, and thus a full match requires six matching alleles.

The string of alleles that is passed from parent to child is known as a haplotype. Because the alleles that determine one's HLA type are located close together on the same chromosome, recombination between these loci is rare. Therefore with rare exceptions, HLA haplotypes remain intact from generation to generation, with each parent randomly selecting one of two possible haplotypes to pass to each of its children. Two siblings will be suitable donors for each other only if each received the same haplotype from each parent. This happens with probability one-fourth. About 70% of all patients needing transplants have no matching sibling. Those without a sibling donor must seek a match from the population at large.

To facilitate non-sibling matches, the developed nations of the world have set up national volunteer registries. Registrants affirm that they will contribute stem cells to any needy patient for whom they are the best available match. Each registrant's HLA type is determined by molecular testing and is recorded along with contact information. The largest of these registries is the National

¹These loci are known as HLA-A, HLA-B, and HLA-DRB. Recent research indicates that success probabilities are improved if there is also a match at the locus HLA-C .

Marrow Donor Program (NMDP) which includes more than 7 million registrants from the United States, Germany, Scandinavia, the Netherlands, and Israel.

Despite the impressive size of the world registry, matches for needy individuals are not always found. The distribution of HLA types is extremely diffuse. There are more than 10 million possible types. Two individuals are more likely to match if they are of the same racial background. Approximately half of the Americans of European ancestry belong to types whose frequency is less than one in one hundred thousand, while twenty percent belong to types with frequency of less than one in a million. The distribution of types is even more diffuse for persons of Asian and African ancestry. Because the world's registries have many more persons of European than of Asian or African ancestry, chances of finding a match are significantly smaller for the latter groups.

In a recent paper [2], we estimated the distribution of HLA types among whites, African-American, Asian-American, and Hispanics. We constructed these estimates estimates of the distribution of haplotypes in each race by M. Mori *et al* [13]. Mori's estimates were based on a sample of about 400,000 NMDP registrants. We estimated the probability that individuals of each race would find a matching HLA type in a registry of given size and racial composition. We did an economic benefit-cost analysis of adding registrants to the NMDP registry and concluded that the expected present value of adding registrants of every race exceeds the cost, with the largest net benefits from adding minority registrants.

Recently, Kollman *et al* [11] have published new estimates of HLA haplotype distributions for each of the races considered by Mori. These estimates are based on a much larger sample of about two million individuals taken from the NMDP registry. Neither Kollman nor Mori were able to make direct estimates of the haplotype distributions of persons of mixed race, since the samples available to them did not contain sufficient numbers of persons known to be of specific racial combinations.

Since a matching donor and recipient must have inherited the same genes from two parents, the best prospect for a person of mixed race is someone of the same mixed-race heritage—but persons of specific mixed-race combinations are relatively scarce. Approximately 2 percent of the population in the United States, 1.5 percent of the population of Canada, and 1.4 percent of the population of the United Kingdom declare themselves to be of mixed race.²

²In the U.S. census of 2000, about 0.28 percent of the population was identified as mixed-race African-American and white and 0.31 percent as mixed-race Asian and white, about 0.8

Although we do not have direct estimates of the HLA type distributions for persons of mixed race, we are able to construct such estimates from estimates of the distribution of haplotypes within single races, by applying the simple combinatorics of sexual diploid reproduction.

In this paper, we estimate the distribution of HLA types for individuals of mixed race and calculate probabilities that persons of each specified mixed-race combination will find a match in the current NMDP registry. We also update our earlier estimates of matching probabilities for single races, using the haplotype distributions reported in the recent Kollman study [11]. We use these statistics to estimate the expected number of lives saved from adding a person of specified race or mixed-race combination to the registry. Finally, using standard economic estimates of the value of a statistical life, we estimate a dollar value of adding registrants of any specified race and compare these to estimated costs of adding new registrants.

2 Methods

The largest available source of data on the distribution of HLA types is the NMDP registry. The NMDP records the self-reported race and determines the HLA type of each registrant. In the early years of the registry, most volunteers were typed only at two loci, HLA-A and HLA-B. As technology improved, new registrants were typed at the three loci, HLA-A, HLA-B, and HLA-DRB1. In a 1997 paper, M. Mori *et al* [13] used a sample of about 400,000 registrants who had been typed at three loci to estimate the distribution of HLA types in each racial subgroup of the U.S. population. More recently, Kollman *et al* [11] constructed new estimates using a sample of about two million registrants, including 1.2 million European-Americans, 250,000 Asian-Americans, 280,000 African-Americans, and 320,000 Hispanics.³ Although the samples used in the Mori and Kollman studies are large, they are not nearly large enough to provide good direct estimates of the distribution of relatively rare types. Many types HLA-types that are present in the population will not appear in the sample. But the mechanics of diploid genetics make it possible, under reasonable assumptions about mating patterns, to use the observed distribution of phenotypes to

percent were identified as white and “some other race”.

³The Kollman study was able to achieve this larger sample partly because the number of the increased number of persons in the registry who have been typed at all three loci and partly because it applied a more general estimation method that extracts information from data about early registrants who had been typed only at two loci.

construct maximum-likelihood estimates of the distribution of HLA haplotypes for each race. Since the number of possible haplotypes at the three-locus level is “only” of the order of 11,000, the data available from the NMDP registry is sufficient to yield reasonably accurate estimates of haplotype distributions. The Mori and Kollman studies followed this strategy and published tables that report estimates of the frequency distribution of haplotypes in each racial group.

2.1 HLA-Type Distributions By Race and Mixed Race

We use the estimated haplotype distributions for “unmixed” races from the study by Kollman *et al* [11] to estimate the distribution of HLA-types for mixed-race combinations. For our purposes, the HLA-type of an individual is determined by the six alleles found in the three genetic loci, A , B , and DRB . The set of six alleles that an individual possesses is known as her *phenotype*. The test used for HLA-typing identifies the six-allele phenotype of each individual, but does not identify the specific haploid triplets from which the phenotype is assembled. However, knowing the probability distribution of haplotypes for the racial group of each parent, we can estimate the probability distribution of phenotypes for individuals of specific mixed race combinations.

In our earlier paper [2], we showed how to calculate the probability that a person of a specified race will find a match in the registry, given the number of people of each race who are registered. Let R_x be the number of persons of race x in the registry and let p_i^x be the fraction of the population of race x that is of HLA type i . The probability that no type i 's are found among registrants of race x is the probability that no type i 's are selected in R_x random draws from the population of race x . This probability is

$$(1 - p_i^x)^{R_x}. \quad (1)$$

A registry contains no persons of type i if there are no type i 's among registrants of any race. Therefore, when R is the vector of registrants by race, the probability that a person of type i has no match of any race in the registry is

$$p_i^0(R) = \prod_x (1 - p_i^x)^{R_x}. \quad (2)$$

It follows that the probability that a person of race x has no match in the registry is

$$\sum_i p_i^x p_i^0(R). \quad (3)$$

It is straightforward to extend this procedure to estimate the distribution of HLA-types in biracial individuals, given estimates of the haploid distributions in each race. A biracial individual with one parent of race X and one of race Y could acquire the six alleles a_1 , a_2 , b_1 , b_2 , dr_1 , and dr_2 in any one of 8 ways. For example, she could inherit the three alleles a_1 , b_1 , and dr_1 from the parent of race X (in the form of haplotype $a_1b_1dr_1$) and the remaining alleles a_2, b_2 , and dr_2 (in the form of haplotype $a_2b_2dr_2$) from the parent of race Y . Alternatively, she could inherit the haplotype $a_2b_1dr_1$ from the parent of race X and the haplotype $a_1b_2dr_2$ from the parent of race Y . There are a total of 8 such combinations. Since we have the estimated haplotype distribution for each race, we can calculate the probability of each of these eight combinations. Adding these 8 probabilities yields the probability that the individual's HLA-type is given by the alleles a_1 , a_2 , b_1 , b_2 , dr_1 , and dr_2 . We run a computer routine that loops through all possible phenotypes, calculating its probability as the sum of the probability of all haplotype pairs from which it could be assembled.

2.2 Expected Lives Saved and Value of Additional Registrants

In order to estimate the probability that persons of mixed race would find a match in the registry, we need estimates of the number of persons in the registry who are of each mixed race combination. The NMDP reports the numbers of self-designated whites, African-Americans, Asian-Americans, and Hispanics who are in the registry. Not every registrant is able or willing to donate when called upon to do so. The NMDP has provided statistics on the availability rate for donors of each race.⁴ We used the NMDP estimates of availability, we have adjusted the size of the “effective” registry of each race to estimate the number of persons of each race who are registered and will be available if called upon to donate. The NMDP reports that 210,000 registrants are of “multiple race”, but did not record the number of registrants of specific multi-race combinations. In the year 2000 U.S. Census, respondents were asked to specify their race and if they were multi-racial, to specify the races to which they belonged. We make the simplifying assumption that NMDP registrants reporting themselves to be multiracial are biracial and that the proportions of those in the registry who are of each biracial combination are the same as the proportions reported in the Census. These statistics are reported in Figure 1 of our Results section.

⁴These figures are reported in Kollman *et al* [10].

Given our estimates of the number of registrants in each racial group and of the distribution of HLA-types within each group, we are able to estimate the probability that a randomly selected person from each group would find no match in the NMDP registry. These results are reported in Table 2.

In our earlier paper [2], we showed how to estimate the expected number of lives saved by adding new registrants. The first step in this process is to estimate the probability that a new registrant of given race will provide a match for someone who did not previously have a match in the registry. Let us define $G_{xy}(R)$ as the increase in probability that a person of race y will have a match in the registry that results from adding a registrant of race x , when the vector of registrants by race is R . Let $p_i^0(R)$ be the probability that there is no potential donor of type i in the registry when the vector of registrants is R . The probability that someone of race y is of type i and has no match in the registry is $p_i^y p_i^0(R)$, and the probability that a new registrant of race x is of type i is p_i^x . Therefore the probability that a person of race y is of type i , has no match in the current registry, and will have a match if an additional person of race x is added to the registry is $p_i^x p_i^y p_i^0(R)$. Summing these probabilities over the types, we have

$$G_{xy}(R) = \sum_i p_i^y p_i^x p_i^0(R). \quad (4)$$

Adding one more HLA type to the registry will only result in an additional stem cell transplant if a patient of that HLA type is in need of a transplant. To calculate the probability that adding a person of race x to the registry will result in an additional transplant to a person of race y during a given year, we must multiply $G_{xy}(R)$ by the number of persons of race y who will seek transplants during that year. We display the resulting estimates in Table 6 of our Results section.

3 Results

The first column of Table 1 reports our estimates of the number of persons of each race and biracial group who are registered by the NMDP. The second column reports estimates of the fraction of registrants who would be available if asked to donate. For the single-race groups, these numbers are based on experience of the NMDP as reported in [10]. For the biracial groups, we have used the mean of the proportions for the two single-race groups involved. The third column is the product of the first and second columns.

Table 1: Number of Registrants and Available Registrants by Race and Biracial Group

Racial Group	Number of NMDP Registrants	Probability Registrant is Available	Number of Effective Registrants
White	5,041,000	0.65	3,277,000
African-American	535,000	0.34	182,000
Asian-American	527,000	0.44	232,000
Hispanic	690,000	0.47	324,000
African-American, White	36,700	0.50	18,200
Asian-American, White	42,700	0.55	23,300
Hispanic, White	77,700	0.56	43,500
African-American, Asian-Amer	6,700	0.39	2,600
African-American, Hispanic	37,400	0.41	15,200
Asian-American, Hispanic	8,700	0.46	4,000

Table 2 reports the probability that a randomly selected member of each racial group will fail to find a willing and able donor in the NMDP registry. These probabilities are calculated using the effective registry sizes estimated in Table 1. It is interesting to see that biracial individuals with one white parent are more likely to find a match in the registry than someone with two parents of the minority race, though less likely to find a match than someone with two white parents. In contrast, children with one African-American parent and one Asian-American parent or with one Asian-American and one Hispanic parent are less likely to find a match than a random person of either parental race.

Our estimates allow us to calculate the probability that biracial patients and those of minority races will find a match among registrants of their own minority group and among white registrants as well as the probability that they will find at least one match in the registry. From Table 3, we see that biracial individuals with one white parent are more likely to find a match than persons who are entirely of their minority ancestry. Although the distributions of HLA haplotypes differ across races, there is substantial overlap. Because of this overlap, and because the number of white registrants in the NMDP is much greater than that of minority groups, biracial individuals with one white parent are more likely to find a match in the registry than those with two parents of their minority race.

Table 2: Probability of No Match by Race or Biracial Group

Race or Biracial Group	Probability of No Match in NMDP Registry
White	0.07
African-American	0.41
Asian-American	0.22
Hispanic	0.17
African-American, White	0.28
Asian-American, White	0.20
Hispanic, White	0.13
African-American, Asian-American	0.49
African-American, Hispanic	0.34
Asian-American, Hispanic	0.27

Table 3: Probability of Match in Registry by Race of Donor and Patient

Race of Patient	Probability of Match from Own Minority	Probability of Match from White Registry	Probability of Having Some Match
Asian	.64	.45	.78
Biracial, Asian-White	.40	.69	.80
African	.41	.30	.59
Biracial, African-White	.36	.59	.72
Hispanic	.63	.69	.83
Biracial, Hispanic-White	.68	.82	.87

Given the distribution of HLA types for biracial individuals, it is straightforward to calculate HLA-type distributions for persons of very complex mixed racial background. This calculation is simplified by the fact that the HLA alleles carried by any individual are inherited from just two grandparents. Consider, for example, an individual whose paternal grandparents are of races W and X and whose maternal grandparents are of races Y and Z . This individual's HLA type will be inherited from one paternal and one maternal grandparent. His HLA type is equally likely to be determined by random draws from each member of one of the four pairs W, Y , W, Z , X, Y , and X, Z . This individual's HLA type distribution is therefore a mixture distribution where the probability

of any given phenotype is an equally weighted average of the probability that this phenotype is found in biracial individuals of types $W, Y, W, Z, X, Y,$ and X, Z .

Consider the special case of marriages between two biracial individuals with the same two racial backgrounds. If both parents have one parent of race X and one of race Y , then each of their offspring will inherit two haplotypes from race X ancestors with probability of $1/4$, two haplotypes from race Y with probability $1/4$, and one haplotype of race X and one of race Y , with probability $1/2$. The distribution of HLA-types for these offspring is a mixture distribution with a weight of $1/4$ on each of the distributions for races X and Y and with a weight of $1/2$ on the distribution for biracial individuals of races X and Y . The probabilities of finding no match in the registry for offspring of such pairs is the corresponding weighted average of the probabilities found in Table 2. The probability that an offspring of two biracial parents, both of whom have one white and one African-American parent, is $\frac{1}{4} \times .07 + \frac{1}{4} \times .41 + \frac{1}{2} \times .28 = .26$. By comparison, the probability for someone with one white and one African-American parent is .28. Similar computations can be made for all possible pairs of biracial parents. Our estimates indicate that for each pair of races, children with two biracial parents have a slightly better chance of finding a match in the registry than children with one parent of each race.

Consider an individual who has one parent of race X and one parent who is biracial with one parent of race X and one of race Y . He will then have three grandparents of race X and one of race Y . With probability $1/2$ he will inherit haplotypes from two grandparents of race X and with probability $1/2$, he will inherit one haplotype each from grandparents of race X and Y . His HLA type distribution will therefore be a mixture distribution with a probability weight of $1/2$ on the distribution for persons of race X and a probability weight of $1/2$ on the distribution for biracial individuals with parents of races X and Y . Table 4 reports the probabilities of finding a match for persons who have three grandparents of the race listed in each row and one grandparent of the race listed in each column.

Adding a person to the registry will result in an additional transplant only if 1) the new registrant's HLA type is not a duplicated by that of another registrant and 2) a patient of this type arrives needing a transplant. The probability that a new registrant of specified race adds a new HLA type to the registry is the same as the probability that a person of that race would be unable to find a match if

Table 4: Probability of No Match for Persons
With Three Grandparents of Row Type, One of Column Type

	White	African-American	Asian-American	Hispanic
White	0.07	0.18	0.14	0.10
African-American	0.35	0.41	0.45	0.38
Asian-American	0.21	0.36	0.22	0.25
Hispanic	0.10	0.26	0.22	0.17

it were needed. Because the registry contains many more whites than persons from minority groups, a new registrant from a minority group is more likely to add a new type to the registry. As Table 2 shows, this probability is only 0.07 for whites and is much larger for the races that are less well represented in the registry.

Although the HLA-type of new registrants from less numerous population groups are more likely to be unique in the group, this by itself does not imply that adding a registrant from a minority group will result in an additional match with a needy patient. Whether this is the case, depends in part on the expected number of patients of their own race who seek transplants. Table 5 reports our estimates of the annual number of patients of each race and biracial classification who sought transplants in 2007.

Table 5: Annual Number of Patients Seeking Transplants by Race

Race	Number
White	3,009
African-Amer	338
Asian-Amer	164
Hispanic	369
Biracial, African-Amer, White	9
Biracial, Asian-Amer, White	14
Biracial, Hispanic, White	19
Biracial, African-Amer, Asian-Amer	2
Biracial, African-Amer, Hispanic	9
Biracial, Asian-Amer, Hispanic	3

Since the proportion of biracial individuals in the population is small, the

number of biracial patients seeking transplants is also small, as shown in Table 5. If the only beneficiaries from biracial registrants were people of their own biracial category, we might conclude that the expected benefits from adding biracial registrants is small relative to that from adding registrants to the larger single-race groups. Table 6 displays the probability that an additional registrant of each race or biracial background will turn out to be the only available match for some patient of each specified racial category. (The numbers reported are the estimated probabilities time 10^5 .) As Table 6 shows, biracial registrants are quite likely to provide unique matches for persons in the larger single-race classifications.

If we look down the column for white recipients, we see that the largest entry is in the row corresponding to persons of biracial African-American and white ancestry and the second largest is in the row for Asian-American and white biracials. This means that if we were to select a single registrant to add to the current registry with the objective of maximizing the probability that a white person would find a match, the best option is biracial of African-American and white ancestry and the second best option is of Asian-American and white ancestry. There is a plausible explanation. The number of whites in the registry is very large and so the types that are not represented are almost certainly very rare in the white population. Many of these rare types will come from individuals who are classified as white, but carry haplotypes passed down from one or more nonwhite ancestors. For such persons, the biracial population is the most promising potential source of a match.

From Table 6, we see that for all races and biracial groups other than whites, the largest entries in the column are the diagonal elements. This means that adding a new registrant of one's own racial group will do more to improve the probability that one will have a match than adding a new registrant from any other group. For example, adding an additional biracial individual of Asian-American and white ancestry is more than twice as likely to provide a match for a previously unmatched Asian-American than adding a registrant who is entirely Asian.

The first column of Table 7 shows, for each race and biracial category, the probability that adding a registrant of this race will provide a match for some patient *of the same race* who would otherwise have no match. The second column shows the probability that an additional registrant of each race will provide a match for a patient *of any race* who would otherwise have no match.

Table 6: Probability that an Additional Registrant is a Unique Match by Race of Registrant and Race of Recipient (times 10^{-5})

Donor Race	Race of Recipient									
	W	Af	As	H	Af-W	As-W	H-W	Af-As	Af-Hi	As-H
W	3.51	0.35	0.14	0.46	0.01	0.02	0.03	0.00	0.01	0.00
Af	3.14	18.12	0.20	1.97	0.18	0.02	0.06	0.04	0.19	0.01
As	2.63	0.41	5.31	0.72	0.01	0.17	0.03	0.03	0.02	0.04
H	3.73	1.80	0.32	3.85	0.05	0.03	0.09	0.01	0.10	0.02
Af-W	4.85	6.60	0.26	2.00	0.26	0.03	0.09	0.04	0.17	0.01
As-W	4.46	0.55	2.04	0.88	0.02	0.39	0.05	0.03	0.02	0.05
H-W	4.02	1.11	0.26	1.76	0.04	0.03	0.08	0.01	0.04	0.01
Af-As	4.28	5.70	1.75	2.26	0.16	0.19	0.07	0.36	0.18	0.08
Af-H	4.20	7.50	0.33	4.27	0.18	0.03	0.10	0.05	0.35	0.02
As-H	4.19	1.19	2.43	2.18	0.04	0.27	0.08	0.07	0.05	0.12

We see that the beneficiaries of additional biracial registrants are more likely to come from the single race populations rather than the population of the same biracial background.

Table 7: Probability that an Additional Registrant is A Unique Match by Race or Biracial Ancestry of Registrant (times 10^{-5})

Race of Registrant	Probability of Unique Match	
	To Own Race	To Any Race
White	3.5	4.6
African-American	18.1	23.9
Asian-American	5.3	9.4
Hispanic	3.9	10.0
Biracial, African-American, White	0.3	14.3
Biracial, Asian-American, White	0.4	8.5
Biracial, Hispanic, White	0.1	7.4
Biracial, African-American, Asian-American	0.4	15.0
Biracial, African-American, Hispanic	0.4	17.0
Biracial, Asian-American, Hispanic	0.1	10.6

A transplant will only “save a life” if the patient is restored to health after the transplant and if this patient would not have survived in the absence of a transplant. Thus to estimate the probability that adding an additional registrant will save a life during a single year, we must multiply the probability that this registrant would be the only registered donor available to some patient by the probability

that this patient will survive if given the transplant and would not survive if the transplant is not available. In [2] we estimated the probability that a transplant would save a life to be approximately 0.21.

It is possible to place a rough money value on adding an individual to the NMDP registry. To do so, we require stronger assumptions than our previous results, but we believe these estimates give a useful notion of the orders of magnitude involved. To make this calculation, we assign a dollar value to the annual flow of expected lives saved and we need to account for the fact that a new registrant is likely to remain in the registry for several years. To place a dollar value on the annual flow of benefits, we multiply the annual expected number of lives saved by a “value of a statistical life.” The value of a statistical life is an estimate of the average amount that an individual is willing to pay for a small increment in survival probability. Estimates of the value of a statistical life vary considerably across studies, but according to a survey by Viscusi and Aldi, they are mainly concentrated in the range from four to nine million U.S. dollars. We assume a value of statistical life of \$6.5 million, the midpoint of this range. This is consistent with the policies of the U.S. Environmental Protection Agency, as reported in their publication “Guidelines for Preparing Economic Analyses” [22], which recommends a VSL equivalent to 6.75 million 2004 dollars.

Persons who join the bone marrow registry can remain in the registry until they reach the age of 61. The mean age of new registrants as reported by the NMDP is 35 years. We will assume that a new registrant will, on average, remain in the registry for 25 years. Although medical technology is bound to change over the next 25 years and the number of persons annually seeking transplants may change dramatically,⁵ we will assume that the annual number of persons of each race seeking transplants will remain constant for the next 25 years and that the probability that a transplant saves the patient’s life will remain constant as well. Following standard practice in economic benefit cost analysis, we will discount future benefit flows by two percent per year. Having made these assumptions, we can impute expected present values for adding persons of each race to the NMDP registry (rounded to the nearest \$100).

⁵The annual number of stem cell transplants from NMDP-registered donors has increased steadily over the decade from 1998 to 2008 at an average annual rate of 8.5 percent.

Table 8: Present Value of Adding a New Registrant:
by Race or Biracial Ancestry of Registrant

Race	Present Value
White	\$1,200
African-American	\$6,400
Asian-American	\$2,500
Hispanic	\$2,700
Biracial, African-American, White	\$3,800
Biracial, Asian-American, White	\$2,300
Biracial, Hispanic, White	\$2,000
Biracial, African-American, Asian-American	\$4,000
Biracial, African-American, Hispanic	\$4,500
Biracial, Asian-American, Hispanic	\$2,800

4 Discussion

4.1 The concept of race

The racial categories, white, African-American, Asian-American, and Hispanic into which NMDP registrants are sorted is coarse and somewhat arbitrary. Since the recorded race of a registrant is self-declared, it indicates a social construction that does not necessarily correspond to genetic inheritance. Statistics show, however that the distribution of HLA types differs markedly between races. For example, the probability that a randomly selected white American will match another randomly selected white is 34 times that of matching a random Asian-American, 16 times that of matching a random African-American, and 6 times that of matching a random Hispanic. These distributional differences have important implications for recruitment of registrants from racial minorities.

Our statistical measurements are built on the Kollman *et al* [11] estimates of haploid distributions within each race. Kollman’s estimates, like those in the earlier study by Mori *et al* [13], are founded on a model that makes two critical assumptions about marriage patterns. The first assumption is that each racial group is *endogamous*, that is marriage occurs almost entirely within races. The second assumption is that conditional on marrying within group, the probability that two people marry is independent of their HLA types.

Since the social construct of race is more likely to influence marriage patterns than genetic classification, the use of self-declared race to determine categories

seems appropriate for the model that is being estimated. Jacobs and Labov [6] collected data on all married heads of households and their spouses from a 1 percent sample of the 1990 U.S. Census. They determined the self-declared race or national origin of each member of each couple. They found that almost 98 percent of marriages of whites and 96 percent of marriages of African-Americans were endogamous. The Jacobs-Labov study shows that approximately 85 percent of Asian-Americans are married to other Asian-Americans and 77 percent of Hispanics are married to other Hispanics.⁶⁷ The genetic composition of the current population depends, of course, on the marriage patterns of their parents' generation, not on current marriage patterns. There is good reason to believe that the current population of Asian-Americans and of Hispanics are children of more endogamous populations than is indicated by current marriages. About 2/3 of the existing population of Asian-Americans were born in Asia and their ancestors for many generations would have had little exposure to non-Asians. About 1/3 of the existing population of Hispanics are immigrants from regions where the population is almost entirely Hispanic.

While the assumption that Asian-Americans marry endogenously is not wildly inaccurate, the assumption that marriage among Asian-Americans is random with respect to HLA type is quite clearly violated. The marriage patterns of the parents of the current generation of Asian-Americans were far from random. Two-thirds of the current population of Asian-Americans are immigrants, coming from several distinct Asian populations that have been geographically separated for many generations. Even after reaching the United States, Asian-Americans have been far more likely to marry within their own nationality than outside of it. Jacobs and Labov [6] find that about 80% of Asian-American marriages are between two people of the same national origin. The distribution of national origins of the current Asian-American population of the U.S. is as follows:

⁶Jacobs and Labov report rates of out-marriage for each of several Asian nationalities. We weighted these rates by the number of marriages of each type to find an average rate of out-marriage. Of those who marry outside of their nationality group, about 40 percent of men and 25 percent of women marry other Asians.

⁷According to Jacobs and Labov, among Hispanics, the marriages of 82 percent of Mexican-Americans, 76 percent of Cuban-Americans and 66 percent of Puerto Ricans were endogenous.

National Origin	Fraction
China & Taiwan	0.24
Indian subcontinent	0.17
Philippines	0.17
Vietnam	0.10
Korea	0.10
Pacific Islander	0.08
Japan	0.07
Other	0.06

Data available from a number of studies indicate that the distribution of HLA types differs significantly among Asian populations [4]. Therefore the fact that Asian marriages tend to be within sub-populations implies that mating is not random with respect to HLA type.

A similar difficulty is found with "Hispanic" as a racial category. The Hispanic population of the United States includes significant subpopulations that differ in ethnic makeup and have had little contact with each other for many generations. About 66 percent of the Hispanic population of the United States is of Mexican extraction, 13 percent come from Central and South America, 9 percent are Puerto Rican, and 4 percent are of Cuban extraction. Genetic admixture studies of Hispanics in the U.S. reveal that Mexican-Americans on average have 30-40 percent Native American ancestry, while immigrants from the Spanish Caribbean have African genetic contributions that range from 20-40 percent and contributions of about 18 percent from the native American Arawaks and Caribs. [8], [9].

These considerations suggest that our procedure for estimating the distribution of HLA phenotypes by reconstruction from estimates of haplotype distributions is likely to be less accurate for the Asian-American and Hispanic groups than for whites and African-Americans. Even if our estimates are reasonably accurate measures of the average distribution of HLA types, much useful information is lost by treating Asian-Americans and Hispanics as aggregate groups. The probability distribution of HLA-types for an individual who is known to be of Japanese ancestry or of Cuban ancestry will certainly be different from the average distribution of Asians or of Hispanics.

Although current rates of intermarriage between African-Americans and whites are low, African-Americans carry a significant amount of genetic material obtained from white ancestors. As Kittles *et al* [8] observes, "The vast majority of contemporary African Americans are descendants of enslaved Africans kidnapped and transported to America during the transatlantic slave trade from

1619 to 1850.” During the period of slavery, there was substantial mixing of the white and African-American gene pool. Kittles *et al* reports that it is estimated that in 1860, “there were 4.5 million people of African descent in the U.S., of which 600,000 were of mixed ancestry or “mulattos”.

Geneticists have developed methods for using genetic markers to estimate *admixture* proportions, that is the proportions of genetic material in a single population that is inherited from members of two or more distinct ancestral populations.⁸ Several studies have estimated admixture proportions from samples of African-Americans. These studies indicate that the percentage of European admixture in the African-American population differs substantially by region, ranging from 3.5 percent in the Gullah sea island community of South Carolina, 10 percent in the rural South, about 20 percent in the industrial North, and 22-35 percent on the West Coast. [8](Figure 2), [14] The admixture of African-American genetic material in the U.S. white population appears to be much smaller.⁹ The geographic differences in the genetic makeup of the African-American population suggests that the accuracy of estimations of HLA-distributions for African Americans could be improved by disaggregating according to region of birth.

We believe estimation of HLA-distributions for more finely distinguished population groups is feasible with existing data and that it will be valuable in directing recruitment by the world’s stem cell registries.¹⁰ A similar problem is addressed by Pritchard, Stephens, and Donnelly [16] “cryptic” population structure. The authors assume that mating is random within sub-populations, but only aggregate distributional data is observed. The authors propose a Bayesian statistical method for extracting from aggregated data on genotypes, the distribution of alleles in sub-populations and the proportions of the total population belonging to each subpopulation.

The problem studied by Pritchard *et al* differs from the one that we pose in some ways. They study the distribution of alleles at several unlinked loci, while Mori and Kollman work with observations of six alleles-phenotypes with the alleles located in three tightly linked, but extremely polymorphic, loci. We have some additional information that can inform our priors. For example, we

⁸See [21] for a brief discussion of these methods and further references.

⁹We have not found results based on large and diverse samples of U.S. whites. Based on a sample of 187 individuals of European-American ancestry living in State College, Pa, Shriver *et al* [21] reports estimates a mean admixture rate of less than one percent.

¹⁰Of course even better estimates would be possible if registries were to collect more detailed information about the ancestral background of registrants.

know the approximate proportions of the Asian-American population who belong to distinct groups, as stated in the table above. These could be used to set prior probability distributions on the proportions of the registry who belong to each group. There also exist several studies that estimate the distribution of HLA-haplotype distributions in relatively small samples from localized regions. Examples include estimates for Korea [7], China [20] and [19], Taiwan [18]) and India [1] for several regions of Europe [3], for France [15], for Sardinia [12] and for Wales [5]. The Proceedings of the Twelfth International Histocompatibility Workshop has published lists of estimated two and three locus HLA haplotype distributions for quite narrowly-defined national and regional populations throughout the world [4]. These estimates are usually based on samples of only a few hundred individuals and hence can provide reliable information only about the distributions of the most common haplotypes. However, it is possible that they could be used in conjunction with large sample phenotype distributions such as the NMDP or other large national bone marrow registries to estimate the distribution of haplotypes in localized populations. The intuition here is that if one observes a phenotype that is likely to be the union of haplotypes h and h' where h is known to be common in population X and not elsewhere, then this is a hint that haplotype h' is more common in population X than elsewhere.

It is likely that the Mori-Kollman style estimates of haplotype distributions for coarse grained populations could be improved by means of an approach similar to that of Pritchard *et al* to estimate haplotype distributions for the major subpopulations of the Asian and Hispanic groups. Similar methods are likely to be useful for sorting by geographic origin within all ethnic groups.

References

- [1] Suraksha Agrawal, K. Arundhati, Uddalak Bhardwaj, and Suhasini Bhatnagar. HLA antigen and haplotype frequency in Bhargavas and Chaturvedies of UP (India). *Indian Journal of Human Genetics*, 5(1):25–30, 1999.
- [2] Theodore C. Bergstrom, Rodney Garratt, and Damien Sheehan-Connor. One chance in a million: Altruism and the bone marrow registry. to appear in *American Economic Review* currently available at <http://repositories.cdlib.org/ucsbecon/dwp/3-07/>, September 2007.

- [3] J G Bodmer, L J Kennedy, J Lindsay, and A M Wasik. Applications of serology and the ethnic distribution of three locus HLA haplotypes. *British Medical Bulletin*, 43(1):94–121, 1987.
- [4] Dominique Charron. *Proceedings of the twelfth international histocompatibility workshop*. EDK, Paris, 1997.
- [5] C. Darke, M.G. Guttridge, J. Thompson, S. McNamara, J. Street, and M. Thomas. Hla class i (A, B) and ii (DR, DQ) gene and haplotype frequencies in blood donors from Wales. *Experimental and Clinical Immunogenetics*, 15(2):69–83, 1998.
- [6] Jerry A. Jacobs and Teresa G. Labov. Gender differentials among sixteen race and ethnic groups. *Sociological Forum*, 17(4):621–646, December 2002.
- [7] Se Jong Kim, In Hong Choi, and Jeu Deuk Kim. HLA-DR antigens and hla-b:dr haplotypes in Koreans. *Yonsei Medical Journal*, 24(1):34–37, 1983.
- [8] Rick A. Kittles, Eunice R. Santos, S. Oji-Nijkeda Nertiti, and Carolina Bonilla. Race, skin color and genetic ancestry: Implications for biomedical research on health disparities. *California Journal of Health Promotion*, 5:9–23, 2007.
- [9] Yann C. Klimentidis, Geoffrey F. Miller, and Mark D. Shriver. Genetic admixture, self-reported ethnicity, self-estimated admixture, and skin pigmentation among hispanics and native americans. *American Journal of Physical Anthropology*, 138(4):375–383, April 2008.
- [10] C. Kollman, E. Abella, R.L. Baitty, P.G. Beatty, R. Chakraborty, C.L. Christiansen, R.J. Hartzman, C.K. Hurly, E. Milford, J.A. Nyman, T.J. Smith, G.E. Switzer, R.K. Wada, and M. Setterholm. Assessment of optimal size and composition of the U.S. national registry of hematopoietic stem cell donors. *Transplantation*, 78(1):89–95, July 2004.
- [11] Craig Kollman, Martin Maiers, Loren Gragert, Carlheinz Müller, Michele Setterholm, Machteld Oudshoorn, and Carolyn Katovich Hurley. Estimation of HLA-A -B -DRB1 haplotype frequencies using mixed resolution data from a national registry with selective retyping of volunteers. *Human Immunology*, 68(12):950–958, December 2007.

- [12] Rosanna Lampis, Laura Morelli, Mauro Congia, Maria Doloretta Macisand Annapaola Mulargia, Miriam Loddo, Stefano De Virgiliis, Maria Giovanna Marrosu, John A. Todd, and Francesco Cucca. The inter-regional distribution of HLA class ii haplotypes indicates the suitability of the Sardinian population for casecontrol association studies in complex diseases. *Human Molecular Genetics*, 9(20):2959–2965, 2000.
- [13] M Mori, P.G. Beatty, M. Graves, KM Boucher, and F.L. Milford. HLA gene and haplotype frequencies in the North American population: the National Marrow Donor Program Donor Registry. *Transplantation*, 64:1017–1027, 1997.
- [14] Esteban J. Parra, Amy Marcini, Joshua Akey, Jeremy Martinson, Mark A. Batzer, Richard Cooper, Terrence Forrester, David Allison, Ranjan Deka, Robert E. Ferrell, and Mark D. Shriver. Estimating African American admixture proportions by the use of population specific alleles. *American Journal of Human Genetics*, 63:1839–1851, 1998.
- [15] B. Pedrona, K. Yakoubenb, D. Adjaoudb, A. Auvrignonc, J. Landmanc, V. Guerina, G. Levergerc, E. Vilmerc, and G. Sterkers. Listing of common HLA alleles and haplotypes based on the study of 356 families residing in the Paris, France, area: Implications for unrelated hematopoietic stem cell donor selection. *Human Immunology*, 66(6):720–730, June 2005.
- [16] Jonathan Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, June 2000.
- [17] Zhenchao Qian. Options: Racial/ethnic identification of children of inter-married couples. *Social Science Quarterly*, 85(3):746–66, September 2004.
- [18] C.K. Shaw, L.L. Chen, A. Lee, and T.D. Lee. Distribution of HLA gene and haplotype frequencies in Taiwan: a comparative study among Min-nan, Hakka, Aboriginal, and Mainland Chinese. *Tissue Antigens*, 53(1):51–64, January 1999.
- [19] Chunmei Shen, Bofeng Zhu, Mengli Liu, and Shengbin L. Genetic polymorphisms at HLA-A, -B, and -DRB1 loci in Han population of Xian City in China. *Croatian Medical Journal*, 49(4):476–482, August 2008.

- [20] Li Shi, Saeko Ogatab, Jian Kun Yua, Jun Ohashic, Liang Yua, Lei Shia, Hao Suna, Keqin Lina, Xiao Qin Huanga and Masaki Matsushitad, Satoshi Horaie, Masaaki Muramatsub, Jia You Chua, and Katsushi Tokunagac. Distribution of hla alleles and haplotypes in jinuo and wa populations in southwest china. *Human Immunology*, 69(1):58–65, January 2008.
- [21] Mark D. Shriver, Estaban J. Parra, Sonia Dos, Carolina Bonilla, Heather Norton, Celina Jovel, Carrie Pfaff, Cecily Jones, Aisha Massac, Neil Cameron, Archie Baron, Tabitha Jackson, George Argyropoulos, Li Jin, Clive J. Hoggart, Paul M. McKeigue, and Rick A. Kittles. Skin pigmentation, biogeographical ancestry and admixture mapping. *Human Genetics*, 112:387–399, 2003.
- [22] U.S. Environmental Protection Agency. Guidelines for preparing economic analyses. Internet publication, 1997. yosemite.epa.gov/ee/epa/eed.nsf/webpages/Guidelines.html.