New reservoirs of HLA alleles: pools of rare variants enhance immune defense

William Klitz¹, Philip Hedrick² and Edward J. Louis³

Highly polymorphic exons of the major histocompatibility complex (MHC, or HLA in humans) encode critical amino acids that bind foreign peptides. Recognition of the peptide–MHC complexes by T cells initiates the adaptive immune response. The particular structure of these exons facilitates gene conversion (GC) events, leading to the generation of new alleles. Estimates for allele creation and loss indicate that more than 10,000 such alleles are circulating at low frequencies in human populations. Empirical sampling has affirmed this expectation. This suggests that the MHC loci have a system for moving valuable and often complex variants into adaptive service. Here, we argue that HLA loci carry many new mutant alleles prepared to assume epidemiologically meaningful roles when called on by selection provoked by exposure to new and evolving pathogens. Because new mutant alleles appear in a population at the lowest possible frequency (i.e., a single copy), they have typically been thought of as having little consequence. However, this large population of rare yet potentially valuable new alleles may contribute to pathogen defense.

Probing MHC polymorphisms

The MHC glycoproteins of vertebrates are expressed on the cell surface and communicate the presence of infectious disease to the host immune system. Within human populations, these HLA loci encode polymorphic glycoproteins carrying dozens of relatively common alleles. These loci on chromosome 6p are made up of two specialized groups. Class I HLA gene products present peptides from intracellular infectious agents and also regulate the activation of natural killer cells. Class II locus product display peptides derived from extracellular infections to regulate the immune response. The impetus for the identification of HLA polymorphisms has been driven by the contribution of these loci to modulating the outcome of tissue transplantation from unrelated donors to patient recipients. Hypervariable exons of HLA loci encode the antigenic recognition site, a pocket in the HLA molecule that binds to pathogenic peptides. The peptide–MHC complex presents a specific stimulus to T cells, thus initiating the immune response. Across human populations worldwide, up to many hundreds of alleles per HLA locus have been identified through the efforts of bone marrow transplant registries. Based on the relative excess of non-synonymous variants, these alleles appear to be under positive selection [1]. Although great progress has been made in understanding the functional biology of HLA polymorphisms, the origin of this variation remains the subject of ongoing debate. Here, we discuss allelic diversity and the genetics of new allele

Glossary

CpG dinucleotide: a pair of adjacent nucleotides that are G and C in a DNA sequence associated with a propensity for CGs in the nearby sequence. That such CpG rich regions have been maintained unaltered within the polymorphic cassettes in the HLA DRB1 locus suggests a possible design for promoting the movement of the polymorphic cassettes in creating new alleles.

Effective allele number (ne): the number of effective alleles is equal to 1 divided by the expected homozygosity (assuming Hardy–Weinberg proportions) at a locus. For example, if there are two alleles with equal frequency, then the expected homozygosity is 0.5 and ne = 2. By contrast, if there is one common allele with a frequency of 0.8 and four rarer alleles with frequencies of 0.05, then the expected homozygosity is 0.65 and ne = 1.54, somewhat less even though there are more actual alleles. This occurs because rare alleles do not contribute very much to expected homozygosity and the effective number of alleles.

Effective population size (Ne): in population genetics, Ne is the number of individuals in an ideal population in which all parents have an equal expectation of contributing to the next generation. Ne is typically smaller than the census size of a population.

Gene conversion (GC): GC is not an exchange of DNA between chromosomal homologs, but rather the donation of a DNA segment from one homolog to the other. This process occurs at a rate some ten times that of crossing-over during meiosis. Intragenic GC events occurring between identical genes are most common, but more rare intergenic GC events between gene family members also occur and play an evolutionary role in mixing more dissimilar sequence between otherwise independently evolving loci. Segmental GC refers specifically to the MHC-characteristic transfer of several polymorphic sites at once, with the possibility of creating a functionally new allele.

Haplotype: a piece of contiguous chromosome, often a heritable unit, useful for visualizing evolutionary processes involving segments of DNA sequence.

Recombination: a primary process of the reduction division of meiosis in which chromosomal homologs pair and exchange portions of one or more arms. Recombination is initiated by a double-stranded break in the DNA of a chromosome during meiosis. A secondary consequence of this initiating event is gene conversion.

Sibship size: the number of offspring in a nuclear family or the surviving births of a single female is an important parameter for determining new allele survival. For example, if a family only has one child, such families lose one half of new mutations by chance alone in the generation following their appearance.

Single nucleotide polymorphism (SNP): the discovery, typing, and analysis of these single genetic variants, describing the variation at a single nucleotide site, are the mainstay of genome-wide studies of genetic variation. It is argued here that, because of MHC functional complexity in the nature of its expressed variation, SNP studies are poorly suited to identify many evolutionarily consequential events responsible for pathogen defense in vertebrates within the MHC.

Variable cassettes: observed structure of the highly polymorphic exons of MHC genes that encode combinations of polymorphic amino acids that bind foreign polypeptides. These comprise the functional unit of GC transfers in the creation of the pattern of protein evolution that is evident in the protein sequences of an MHC locus.
creation, which we believe reveal a previously unrealized abundance of genetic variation in our species.

Dynamics of MHC gene conversion
In eukaryotes, a double-stranded break in the DNA during early meiosis is the crucial initiating event for transferring information between chromosomal homologs. These breaks are resolved in two ways [2]. First, a crossover event may occur involving the exchange of interhomolog arms, thereby ensuring chromosomal segregation and resulting recombination. Second, the resolution of the double-stranded break can result in non-crossover repair. Both of these modes can include GC (see Glossary), in which a segment of DNA is non-reciprocally transferred from one chromosomal homolog to the other. Because non-crossovers occur more frequently than do crossovers, most GC events are not associated with a crossover. Several features of GC events suggest their involvement in generating HLA allelic diversity, including the short length of transferred segments that they produce [3], their tendency to be concentrated in regions of high CpG dinucleotide content, which is a hallmark of HLA genetics [4,5], higher frequency of occurrence compared with crossover events [2,3], and detectability only in regions having sequence variation.

Point mutations initially set the stage for allelic diversification via GC, especially in multi-gene families [6]. GC tracts can be significantly shorter than gene size, which from both theoretical [7] and observational [6,8] evidence generates new haplotypes by segmental GCs. Empirical studies have identified many candidates of past information transfer among gene family members of the MHC [9–16]. Taken together, this evidence indicates that GCs have created extensive variation at MHC loci by shuffling small segments of DNA between MHC alleles.

We propose that this mechanism is ongoing and serves to create new HLA alleles within the population by transferring immunologically meaningful variation. The DR beta 1 (DRB1) second exon, for example, is a patchwork of DNA segments of two types: one set is highly polymorphic with nine or more cassettes of sequences, each made up of one to several codons. These cassettes can have up to ten variant sites, each influencing antigen recognition in the DR molecule. In stark contrast, the intervening regions consist of entirely conserved codons with high CpG dinucleotide content. Tellingly, CpG dinucleotides are associated with recombination and GC hot spots [5]. This combination of elements leads to expanded pathogen-binding capacity, allowing the vertebrate host to deal with rapidly evolving microbial pathogens (Box 1). A single GC event moving an existing and adaptively proven polymorphic cassette into a new exonic sequence has a high likelihood of creating a new allele capable of novel antigen-binding characteristics. Thus, new alleles appear pre-adapted for presentation of an expanded repertoire of pathogen-derived peptides.

A total of 5251 formally named HLA alleles across eight genes have a frequency of less than 0.005, and most of those are even more rare [17]. The number of frequent alleles for a given population is smaller, with the common alleles ranging from approximately 12 for DQ beta 1 (DQB1) to a few dozen for HLA-B. In a sample of 6000 high-resolution-typed European American samples, the total frequency of those rare alleles as a group was less than 1% across the eight loci, ranging from 0.0% for HLA DQB1 to 0.7% for HLA B [18]. Because of extensive allele sharing across related populations, we predict that a complete tally of HLA alleles will comprise mostly new alleles, including both new and recently created mutants undergoing the gradual random walk to extinction. Because of improved typing quality and larger sample sizes, it is now feasible to detect this new allelic class, despite their rarity and the fact that the population lifetime of these new alleles may be limited.

New mutants: predicting numbers through effective population size and mutation rate
As a primary creator of new MHC alleles [19], the GC rate is an important parameter. Two studies recording GC rates in the MHC examined spermatogenesis in heterozygous males. Using an MHC class II model of intergenic conversion (GC between loci), a rate of 0.000025 per gamete per generation was estimated for mice [20]. In a second study, experimental data on intragenic GC at the HLA DPB1 locus yielded estimates of approximately 0.0001 events per gamete per generation [21]. It should be noted that, although intergenic GC and de novo point mutations are several times less frequent compared with intragenic GCs [22,23], these other processes are still thought to make important contributions to MHC allelic variation.

Estimates of human long-term effective population size \(N_e\) are also important for determining the amount of variation. Based on anthropoid ape–human divergence [24,25], \(N_e\) is estimated to be approximately 10 000. As a whole, the human population is now seven billion. However, humans are not panmictic, and each couple does not produce the same number of offspring; both of these factors influence \(N_e\). HLA history can be viewed as occurring in two phases: the long evolutionary history of approximately stable and fairly small population levels, including many rural populations up to the present, and the more recent generations in which humans have become urbanized and mobile. As recently as 1800, only 6% of people were urban dwellers, but this had rapidly increased to 50% by 2010 (http://www.prb.org/Educators/TeachersGuides/Human-Population/Urbanization.aspx). Even though the existing \(N_e\) is much greater (e.g., \(N_e\) for European populations is estimated at 1.1 million [26,27]), the current mutational composition of our species is still influenced by the long history of smaller \(N_e\) values.

Using a range of estimates of \(N_e\) values, the mutation rate and the frequency of a newly created allele and assuming no selective difference between alleles (neutrality), we can determine reasonable values for a given gene of the number of new alleles per generation, the effective number of alleles present at one time in human populations, and the mean persistence (in generations) of those alleles (Box 2). This reveals that new alleles appear at a rate ranging from 0.3 to as much as 600 per generation, and that between 1.2 and 401 effective alleles are present in a given population, depending on its size. The actual number of alleles in a population may be much greater than the
Box 1. New MHC alleles: GC versus point mutations and recombination.

**GC** is a ubiquitous process and frequent contributor to mutational evolution, and we describe its role here in more detail. A possible GC event occurring in DRB1 exon 2 in a heterozygous parent with chromosomal homologs A and B is illustrated in Table I. The DRB1 structure of exon 2, including the monomorphic CpG-rich segments ‘—’ and additional hypervariable segments from each homolog ‘XXXX’ and ‘YYYY’ are indicated. In the example in Table I, we see that the polymorphic LLEQ motif (in single-letter amino acids) is moved from A into B to create a new allele.

However, it must be kept in mind that GC, point mutations, and recombination have each competed for attention as participants in the creation of new alleles in the MHC [10,11,14,16]. Although these findings argue that GC may be primarily responsible for variation at the MHC, it is clear that point mutations and recombination have also been important.

Better understanding of the functional distinctions between the class I and class II MHC alleles may point to different mechanisms or emphases from locus to locus. For example, HLA B is the most polymorphic locus, so it would be interesting to know whether this is due to higher new allele generation rates compared with other class I loci, or to subsequent selection favoring greater allele diversity. Similarly, we would like to know whether distinctive mutation rates or selection at the DQ heterodimers underlie the uniform allele frequencies seen at these loci across human populations. HLA-C alleles have been something of a conundrum, demonstrating both antigen presentation to T cells and a central role in the regulation of natural killer cells (KIR activation and inhibition). It has been shown that HLA C variation has a history rooted in differential gene expression due to the transfer and subsequent maintenance of a downstream regulatory segment from HLA-B some 3 million years ago [59]. So-called group 1 and group 2 HLA-C alleles neatly conform to the presence and absence, respectively, of the HLA C derived segment. A program to further probe the origin of HLA alleles would include new experiments to further uncover new HLA creation during meiosis by examining gametes in heterozygous donors in class II and class I loci with sample sizes sufficient to estimate frequencies in at least the 0.0001 range. All identified new alleles should include typing on the parents of a proband to identify new mutants. The tally of events for each locus by type of event (intra/inter-genic GC, point mutation, recombination, and other) could prove very informative.

<table>
<thead>
<tr>
<th>Table I. GC event creating a new allele in the DRB1 exon 2.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homolog A (donating strand):</td>
</tr>
<tr>
<td>Homolog B (recipient strand):</td>
</tr>
<tr>
<td>Resulting new allele:</td>
</tr>
<tr>
<td>XXXX            LLEQ             XXXX</td>
</tr>
<tr>
<td>YYYY            ILED             YYYY</td>
</tr>
<tr>
<td>YYYY            LLEQ             YYYY</td>
</tr>
</tbody>
</table>

Effective number of alleles because rare alleles do not contribute much to this measure. We expect the persistence of alleles to range from five to ten generations over the \( N_e \) size range. These calculations clearly demonstrate the potential existence of a huge reservoir of extremely rare alleles.

The expected equilibrium number of HLA alleles and the number of new mutants at a given locus currently present in the total human population can also be determined. The urban and rural populations are considered separately, each comprising approximately 50% of the total population of seven billion (http://www.prb.org/Educators/TeachersGuides/HumanPopulation/Urbanization.aspx). Taking \( N_e \) (rural) as 5000 and \( N_e \) (urban) as 1 million along with the calculations of Box 2, the effective number of alleles per locus will be 2.1 million and 1.4 million, respectively. Thus, we calculate that as many as 3.5 million rare alleles may be present at a given HLA locus, assuming that \( N_e = N \) such that there are 70,000 rural populations and approximately 3500 urban populations. Using the expected number of alleles for these \( N_e \) from Box 2, the total number of expected alleles per locus at equilibrium is approximately 3.5 million. Assuming a mutation rate of 1/10 000 and 14 billion parental gametes, the total number of new mutant HLA alleles per locus in the current generation of humans is some 1.4 million.

High-throughput and high-resolution HLA typing based on many thousands of individuals has been used to expand bone marrow and stem cell registries [28–31]. This sampling across several ethnicities has revealed a wealth of newly identified alleles, most being rare and often composed of multiple nucleotide changes. These studies offer an initial empirical glimpse of the rare alleles we predict based on our calculations. Future research will be able to document the proportion of these variants that are new and/or derived from GC, as well as provide insight into the proportions that are potentially advantageous or detrimental.

**Durable MHC selection**

Although the ideal variant for following evolution has often been viewed as being under neutral selection, the MHC is useful for evolutionary studies, despite a long history of selection. Many parts of the selective puzzle surrounding MHC evolution are being assembled: experimental studies have shown only populations bearing MHC-resistant alleles survive specific pathogen exposures [32,33], bioinformatics has confirmed that allelic divergence coincides with differences in epitope binding for DRB1 alleles [34], observations regarding the human and chimpanzee MHC A and B locus differences in allelic binding that were estimated for a large suite of viral peptides are explainable through the distinct histories of the two species [35], and comparative studies demonstrate a role for convergent evolution in creating the specific MHC allele composition of a species [36,37]. Especially pertinent to our discussion here, the action of GC in MHC diversification following population expansion in a passerine bird has also been documented [38]. MHC selection is stabilizing, in that it maintains high levels of heterozygosity involving many alleles. The evenness of HLA allele frequency spectra at HLA loci for common alleles indicates a history of balancing selection at some but not consistently at all loci and population samples [39,40].

Each human population carries an operational set of common alleles at each HLA locus (>0.001, ranging up to a few dozen at HLA B) [18,40]. The population heterozygosity of these common alleles is high, typically exceeding 90%, which is a value of functional significance for a population as heterozygotes permit the presentation of a wider array of peptides than is possible for homozygotes. We suggest
Box 2. Expectations of allele creation and loss

The four parameters of interest to quantify variation here are new alleles per generation and, in a neutral model of gene evolution, \( N_e \), the frequency of a new allele is 1/(2\( N_e \)), and the mutation rate, \( u \). Here, we assume that the mutation rate generating new alleles \( u \) is equal to the GC rate and that \( N \) is the census population size. Armed with estimates of these parameters, we use the population genetics theory concerning the birth–death process of new mutants to give: (i) new alleles per generation; (ii) expectations of probability of allele loss; (iii) the effective number of alleles segregating in a population; and (iv) the average persistence time of new alleles.

(i) Probability of loss of a newly arisen allele in a single generation \([49]\):

\[
P_{\text{loss}} = \left(1 - \frac{1}{N_e}\right)^{N_u} \approx \frac{1}{e}
\]

(ii) New alleles per generation \([49]\):

\[
n = 2Nu
\]

(iii) Number of effective alleles expresses the number of alleles expected to be seen in the population at equilibrium \([50]\):

\[
n_e = 4N_eu + 1
\]

(iv) The mean persistence time in generations of a new allele destined for loss \([51]\):

\[
T = 2 \left(\frac{N_e}{N}\right) \ln(2N)
\]

Using pertinent parameter values, the Table I gives the number of new alleles per generation assuming that \( N \) is approximately \( 3N_e \) \([52]\), the effective number of alleles, and the mean allele persistence for six values of \( N_e \) with \( u \) set to 0.0001. As \( N_e \) increases from 500 to one million, the number of new alleles per generation increases from 0.3 to 600, and the number of effective alleles at equilibrium changes from 1.2 to 401. Over the same range of \( N_e \) allele persistence doubles from 5.3 generations to 10.4. The probability of fixation for a new allele is 1/(2\( N_e \)). For a new allele with a selective advantage of \( s \) in a growing population, it is approximately \( 2(s+1)/r \) \([53]\). If we let \( s \) approach 0, then this becomes approximately \( 2r \). In other words, the probability of fixation becomes a function of the growth rate. For the contemporary human population, this suggests that many new mutants are not lost but instead maintained in the population.

Table I. HLA allele creation rates, effective allele numbers, and allele persistence.

<table>
<thead>
<tr>
<th>( N_e )</th>
<th>( N_u ) description</th>
<th>New alleles per generation assuming that ( N \approx 3N_e )</th>
<th>Effective allele number</th>
<th>Allele persistence in generations</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>Founding populations ([54]); rural Eurasia ([55])</td>
<td>0.3</td>
<td>1.2</td>
<td>5.3</td>
</tr>
<tr>
<td>1000</td>
<td>African–non-African populations ([56]); human populations ‘recent’ ([57])</td>
<td>0.6</td>
<td>1.4</td>
<td>5.8</td>
</tr>
<tr>
<td>5000</td>
<td>Long term from primate comparisons ([25]); human population history ([58])</td>
<td>3.0</td>
<td>3.0</td>
<td>6.7</td>
</tr>
<tr>
<td>10 000</td>
<td>Transition to large urban</td>
<td>6.0</td>
<td>5.0</td>
<td>7.3</td>
</tr>
<tr>
<td>50 000</td>
<td>Contemporary urban</td>
<td>30</td>
<td>21</td>
<td>8.4</td>
</tr>
<tr>
<td>1 000 000</td>
<td></td>
<td>600</td>
<td>401</td>
<td>10.4</td>
</tr>
</tbody>
</table>

Note that most alleles are derived from the birth–death process of new allele evolution. The allele frequency spectrum (Figure 1) conveys an overview of the evolutionary dynamics present in three frequency categories of alleles: common, purgatory, and rare. Common alleles constitute the functional allelic reservoir in a population responsible for the outcome of the encounter of each generation with pathogens. The rare alleles are so uncommon that they are especially subject to stochastic forces, pushing many to extinction, but at the same time offering quality targets for selection in the face of new pathogenic challenges. The alleles in purgatory are in the middle of the frequency

![Figure 1](image_url)

**Figure 1.** The three states of the human leukocyte antigen (HLA) allele frequency spectrum (common, purgatory, and rare groups) as allelic roles evolve through time. The common alleles are held in place by ongoing balancing selection, whereas rare alleles are especially subject to stochastic loss through the life and death process of new mutants \( (e.g., [48], p. 370) \). The rapid utilization of new mutant HLA alleles through selection may have been particularly common at past periods at the expanding edge of populations that moved into new environments. Those alleles in the purgatory zone can either ascend to common allele status or fall into the rare group with probable extinction. When better empirical data on the frequencies of rare HLA alleles and additional estimates of gene conversion rates become available, theoretical approaches \([44]\) can then be used to further evaluate the factors involved and compare HLA variation to variation at other genetic markers.
spectrum, caught between those common and epidemiologically functional alleles and rare alleles on the edge of survival.

The rapid increase in the human population has led to the recognition and empirical identification of a higher than expected incidence of low-frequency variants in the genome as a whole [26,41–43]. New mutant loss in one generation occurs as a function of the sibship size \(s\) in the nuclear family with \(P(\text{loss}) = 0.5^s\). The larger family sizes that come with population growth result in an increasing probability of allele retention. The probability of loss of a single allele (among the two parents) is 0.5 for \(s = 1\), whereas for \(s = 6\), this falls to 0.016. Because European and Asian out of Africa populations have accounted for the most spectacular extent of human population growth in recent generations, they are expected to have the greatest share of excess new alleles compared with Africans [44].

**Number of possible alleles**

The unique binding characteristics of a particular antigenic epitope are encoded by small cassettes of variable amino acid sequences. These variant cassettes are the critical unit of sequence typically transferred during GC. For DRB1, at least, the intervening sequences of the hypervariable cassettes consist of nearly monomorphic sequence with high CpG content [5], conditions that are prime for the action of GCs to shuffle gene segments between homologs during meiosis. If mutations were to arise via GC only through transfer of these cassettes, how many alleles are possible? This can be calculated as the product of the polymorphism present in each cassette at a locus (Table 1). Using the common alleles of Europeans, for HLA-A, -C, and -B, 155, 80, and 194 million alleles, respectively, are possible, and 3.9, 1.2, and 0.2 million alleles are possible for DR, DQ, and DP molecules, respectively. With background analyses in hand [5], DRB1 is the best-documented case for this model.

Some fraction of mutational events in HLA allele creation may result in alleles already present. A study of new allele creation in sperm of heterozygous males observed one such allele recreated at DPB1 [21]. Additional instances based on population data recorded a DRB1\(^*\)08 allele in the Americans that is identical in amino acid sequence to one present in African populations [28]. The DRB1\(^*\)08 allele is a member of an ancient allele lineage also found in gorillas [36,45]. Some DRB1\(^*\)04 alleles have similarly repeatedly arisen in human populations across the globe. Thus, some portion of new alleles created by GC can result in the recreation of existing alleles.

The potential allelic polymorphism through GC by movement of sequence cassettes shown in Table 1 illustrates another aspect of MHC variation that sets it apart from the information available in the now extensive genome-wide studies based on single nucleotide polymorphisms (SNPs). There are important differences between the many SNP-based studies demonstrating an abundance of new rare mutants accumulated due to the rapid growth of expanding human populations, and the similar results seen here, in which HLA variation also shows an excess of rare new mutants. Although the HLA system is also impacted by rapid population growth and the rise of new mutants, the arguments here also imply a level of evolutionary design in the utility of new immediately functional variable cassettes transferred by GC in the creation of new

---

**Table 1. Number of possible HLA alleles based on exchange of DNA sequence cassettes in hyperpolymorphic exons: class I (e2 and e3) and class II (e2)\(^{a,b}\)**

<table>
<thead>
<tr>
<th>HLA locus</th>
<th>No. of alleles(^a)</th>
<th>No. of cassettes</th>
<th>Cassette polymorphism</th>
<th>Possible alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Class I</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-A e2 and e3</td>
<td>23</td>
<td>13</td>
<td>3-2-8-7-4-6-4-2-6-10-5-2-4</td>
<td>154 828 800</td>
</tr>
<tr>
<td>HLA-C e2 and e3</td>
<td>20</td>
<td>16</td>
<td>4-3-2-2-4-2-4-10-6-3-6-4-2-3-2</td>
<td>79 626 240</td>
</tr>
<tr>
<td>HLA-B e2 and e3</td>
<td>38</td>
<td>14</td>
<td>3-3-2-5-8-6-3-8-9-2-3-6-4-4</td>
<td>537 497 120</td>
</tr>
<tr>
<td>Total class I alleles</td>
<td></td>
<td></td>
<td></td>
<td>771 952 160</td>
</tr>
<tr>
<td><strong>Class II</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DRB1 e2</td>
<td>31</td>
<td>9</td>
<td>9-8-6-2-5-10-6-3-5</td>
<td>3 888 000</td>
</tr>
<tr>
<td>DQ(^b)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set 1 (DQA1 e2 and DQB1 e2)</td>
<td>4, 9</td>
<td>6, 7</td>
<td>3-2-3-3-2-3-3-5-3-3-3</td>
<td>1 180 980</td>
</tr>
<tr>
<td>Set 2 (DQA1 e2 and DQB1 e2)</td>
<td>6, 8</td>
<td>3, 7</td>
<td>2-2-2-2-4-2-3-4-2</td>
<td>6144</td>
</tr>
<tr>
<td>DQ total</td>
<td></td>
<td></td>
<td></td>
<td>1 187 124</td>
</tr>
<tr>
<td>DP: DPA1 e2 and DPB1 e2</td>
<td>6, 18</td>
<td>5, 7</td>
<td>2-2-2-2-2, 3-2-5-4-5-3-2</td>
<td>115 200</td>
</tr>
<tr>
<td>Total class II alleles</td>
<td></td>
<td></td>
<td></td>
<td>5 190 324</td>
</tr>
</tbody>
</table>

*This example is based on the variation observed in the common alleles present in Europeans.

\(^a\)NB. We calculate the total number of possible such cassettes at each HLA locus for a single well-sampled group, the Europeans, by calculating the product of the polymorphism of each cassette for the resulting gene product. Based on the consistency of the extent of polymorphism found in other human groups [23], other large exogamous populations will probably show similar patterns. For the three HLA class I molecules, there are two hypervariable exons. For the two class II molecules heterozygous at each of the heterodimers (i.e., DQα and DPβ), distinct antigen-binding molecules derive from a single hypervariable exon at each of the heterozygous members. By contrast, the DR molecule derives its antigen recognition polymorphism from the single beta chain exon. The total possible polymorphism attainable through GC in existing variable cassette polymorphisms of a population can be calculated as the numerical product of the number of variable sites in each polymorphic cassette (that is, for HLA-A, 3 X 2 X 8 X 7... = 154 828 800). An individual polymorphic cassette has from two to ten variants per cassette. Some cassettes consist of single amino acid sites, whereas the most complex cassettes range up to eight amino acids. On this basis, the common European alleles for the six class I and class II HLA loci reveal large numbers of possible alleles. Reflecting distinct functional roles, the total number of alleles is orders of magnitude higher for the class I loci compared with that of class II loci. The combined variation from both alpha and beta chain variability is the primary contributor to this difference. The somewhat analogous variation of both chains (DQα–DQβ) and (DPα–DPβ) in the DQ and DP molecules makes a much smaller contribution due to the intrinsically lower number of polymorphisms per cassette.

\(^b\)European allelic polymorphism from [18,60,61], with alleles having frequencies greater than 0.001.

\(^c\)DQ has two mutually exclusive sets of polymorphic heterodimers.
alleles. This further suggests that SNP analyses are often opaque to the functional significance of multiple complex variants present both among and within the variable cassette system underlying the postulated mechanisms for new HLA allele creation.

Practical consequences
HLA matching in stem cell transplantation for a variety of conditions now results in more than 50,000 transplants each year [46]. The HLA matching of recipients and donors is coordinated worldwide by an agency managing the search for matching pairs from transplant directories [46]. This biomedical activity is largely responsible for current general knowledge of HLA polymorphisms and genetics. Another potential driver of this search has been researcher credit through publication of new allele identification and discovery. For the transplant practitioners dependent on accurate and complete HLA allele definitions, the new allele recording system [47] has been an invaluable boon. Yet, as we discuss here, the completion of the allele identification task is both daunting and pointless. Daunting because of the huge numbers of still unidentified alleles present in the human population today, and pointless, because most of these alleles are present for short amounts of time and in small numbers, making the discovery of unrelated HLA matches extremely unlikely. For a patient bearing a rare allele, the best chance for a match will be found in that patient’s first-degree relatives.

Concluding remarks
The MHC loci have genetic structures in their hyperpolymorphic exons that promote GC events among the variable sequence cassettes within these exons, thus creating functional new alleles. For the HLA complex, GC is a system for creating and moving valuable complex variants into adaptive service. For the loci demonstrating sizeable population reserves of new GCs represented by the numbers of new variants postulated here, including HLA A, C, B, and DRB1, with an estimate of two untested alleles for every 10,000 individuals at each locus, a reserve of new peptide presentation possibilities are waiting in the wings at each locus. Furthermore, we suggest that these variants will be mostly derived from motifs that have proven effective at pathogenic peptide presentation, forming a reservoir for enhanced immune defense.

References
2 Rosu, S. et al. (2011) Robust crossover assurance and regulated interhomolog access maintain meiotic crossover number. Science 334, 1286–1289
5 von Salome, J. and Kukkonen, J.P. (2008) Sequence features of HLA-DRB1 locus define putative basis for gene conversion and point mutations. BMC Genomics 9, 228
9 von Salome, J. et al. (2007) Full-length sequence analysis of the HLA-DRB1 locus suggests a recent origin of alleles. Immunogenetics 59, 261–271
31 Lazaro, A.M. et al. (2012) Increased HLA class I and II diversity as 72 novel alleles are identified in volunteers for the National Marrow Donor Program Registry in 2010. Tissue Antigens 79, 50–57
Opinion

37 Srithayakumar, V. et al. (2012) Evidence for evolutionary convergence at MHC in two broadly distributed mesocarnivores. Immunogenetics 64, 289–301
55 Evisukov, A.N. et al. (1996) [Geography of effective size in the rural population of northern Eurasia: effective size and the intensity of random genetic drift]. Genetika 32, 1396–1405
57 Tenesa, A. et al. (2007) Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17, 520–526