Genotype 1 and global hepatitis C T-cell vaccines designed to optimize coverage of genetic diversity

Karina Yusim,1 William Fischer,1 Hyejin Yoon,1 James Thurmond,1 Paul W. Fenimore,1 Georg Lauer,2 Bette Korber1 and Carla Kuiken1

Correspondence
Karina Yusim
kyusim@lanl.gov

1Los Alamos National Laboratory, Theory Division, Los Alamos, NM 87545, USA
2Gastrointestinal Unit, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

Immunological control of hepatitis C virus (HCV) is possible and is probably mediated by host T-cell responses, but the genetic diversity of the virus poses a major challenge to vaccine development. We considered monovalent and polyvalent candidates for an HCV vaccine, including natural, consensus and synthetic ‘mosaic’ sequence cocktails. Mosaic vaccine reagents were designed using a computational approach first applied to and demonstrated experimentally for human immunodeficiency virus type 1 (HIV-1). Mosaic proteins resemble natural proteins, but are assembled from fragments of natural sequences via a genetic algorithm and optimized to maximize the coverage of potential T-cell epitopes (all 9-mers) found in natural sequences and to minimize the inclusion of rare 9-mers to avoid vaccine-specific responses. Genotype 1-specific and global vaccine cocktails were evaluated. Among vaccine candidates considered, polyvalent mosaic sequences provided the best coverage of both known and potential epitopes and had the fewest rare epitopes. A global vaccine based on conserved proteins across genotypes may be feasible, as a five-antigen mosaic cocktail provided 90, 77 and 70 % coverage of the Core, NS3 and NS4 proteins, respectively; protein coverage diminished with increased protein variability, dropping to 38 % for NS2. For the genotype 1-specific vaccine, the H77 prototype vaccine sequence matched only 50 % of the potential epitopes in the population, whilst a polyprotein three-antigen mosaic cocktail increased potential epitope coverage to 83 %. More than 75 % coverage of all HCV proteins was achieved with a three-antigen mosaic cocktail, suggesting that genotype-specific vaccines could also include the more variable proteins.

INTRODUCTION

An estimated 170 million people are infected with hepatitis C virus (HCV) worldwide (Armstrong et al., 2000; Lauer & Walker, 2001). Acute HCV infection leads to chronic infection in 70–80 % of infected individuals (Alter et al., 1999; Conry-Cantilena et al., 1996) and can result in liver cirrhosis, liver failure and hepatocellular carcinoma (Lauer & Walker, 2001). The current treatment with pegylated alpha interferon and the nucleoside analogue ribavirin has improved clinical outcomes (Davis et al., 1998; Fried et al., 2002; Manns et al., 2001; Zeuzem et al., 2000), but is expensive and requires sophisticated medical management, which is out of reach for the majority of infected people. Development of a prophylactic vaccine to prevent the spread of HCV infection remains paramount.

Several findings indicate that immunological control of HCV is possible. In humans and chimpanzees, spontaneous eradication of HCV can occur during acute infection (Cooper et al., 1999; Gerlach et al., 1999, 2003; Lechner et al., 2000; Thimme et al., 2001). Spontaneously cleared HCV infection reduced the likelihood of developing a chronic infection in continuously exposed drug users (Mehta et al., 2002). Reinfection after initial clearance had a shorter duration and a lower peak viraemia in the chimpanzee model (Shoukry et al., 2003). HCV-specific cytotoxic T-lymphocyte (CTL) and T-helper immune responses in early infection have been shown to correlate with clearance of HCV infection (Cerny & Chisari, 1999; Cooper et al., 1999; Cucchiariini et al., 2000; Darling & Wright, 2004; Erickson et al., 2001; Gruner et al., 2000; Lechner et al., 2000; Ward et al., 2002) and HCV-specific CTLs have been shown to persist for years after infection (Cooper et al., 1999; Darling & Wright, 2004). The fact that HCV can be controlled by the host immune system is a cause for optimism regarding the development of HCV vaccines (Houghton & Abrignani, 2005). Both B cell- and T cell-based prophylactic and therapeutic vaccines are under development (Inchauspe & Michel, 2007; Inchauspe et al., 2009; Stoll-Keller et al., 2009).
One of the major challenges for developing an effective HCV vaccine is its genetic diversity (Houghton & Abrignani, 2005). For a T cell-based vaccine to be effective, responses covering a wide range of variants are required, as single amino acid substitutions can completely abrogate recognition by CD8+ cytotoxic and CD4+ T-helper cells (Erickson et al., 2001). Reverse vaccinology strategies (Rappuoli, 2001; Serruto et al., 2009) may be critical for designing vaccine antigens that can contend with highly variable viruses such as HCV or human immunodeficiency virus (HIV).

For HIV-1, a similarly variable virus, artificial vaccine sequences that minimize differences from the circulating strains, such as consensus sequences or ancestral sequence reconstructions, have been considered (Gaschen et al., 2002). Consensus and ancestral envelope proteins retained folding and conformational antibody-binding characteristics, and responses to a consensus sequence showed enhanced B- and T-cell cross-reactivity compared with natural strains (Doria-Rose et al., 2005; Gao et al., 2005; Santra et al., 2008).

Recently, a new computational approach to the design of polyvalent vaccine antigens for T cell-based vaccines was developed for HIV-1 (Fischer et al., 2007; Korber et al., 2009). These antigens consist of sets of ‘mosaic’ proteins, which are computationally generated recombinants assembled from fragments of natural sequences using a genetic algorithm. Mosaic proteins resemble natural proteins, but are optimized to maximize the coverage of common potential T-cell epitopes (9-mer peptides) found in a population of natural sequences, and to minimize the inclusion of rare epitopes to avoid vaccine-specific responses. Sets of mosaic proteins provide coverage of the most common 9-mers in the circulating population, but enable the delivery of these variants in the form of intact proteins that could be processed naturally and delivered readily in a vaccine. For example, a set of four mosaic proteins approaches inclusion of the four most common variants of every potential epitope in the input population of viral sequences (Fischer et al., 2007; Korber et al., 2009).

HIV-1 mosaic antigens can be synthesized and expressed and are immunogenic in mice (Fischer et al., 2008). In a recent study, two- or three-mosaic sets of HIV-1 envelope proteins optimizing 9-mer coverage of a set of diverse global HIV sequences were compared with sets of two or three natural-strain proteins selected to provide, in combination, the best 9-mer coverage of HIV-1 sequences (Kong et al., 2009). Mosaic sets elicited multiple CD4+ and CD8+ responses, and the breadth of responses was many-fold higher than the responses to natural-strain antigens (Kong et al., 2009) when assayed using peptides that contain the most common epitope variants found in the sampled global diversity (Li et al., 2006). Similarly promising results were obtained in recently completed studies in macaques: HIV-1 mosaic antigens elicited responses with greater breadth (more responses to diverse variants overall) and greater depth (more variants recognized per response) than did natural strains (Barouch et al., 2010; Santra et al., 2010).

HCV vaccinology faces even greater population diversity challenges than HIV vaccinology, and the underlying phylogenetic structure at the population level is different. In this study, we assembled representative global HCV sequences, assessed the phylogenetic structure with vaccine considerations in mind, and then proposed an optimized solution to the diversity issue for both genotype 1 and global HCV T-cell vaccines. Modifications of the mosaic strategies developed for HCV were employed, including serial addition of distinct genotypes, to achieve reasonable levels of HCV global coverage.

**RESULTS**

**Comparing HCV and HIV-1 variability**

An important consideration for any HCV vaccine is its scope. Whilst genotype 1 is prevalent in Europe and North America, non-1 genotypes are common in many areas of high HCV prevalence (Fig. 1a). Genotype 3 is common in the former USSR; genotype 4 is common in Africa; genotypes 3, 4 and 5 are common in the Middle East; and genotypes 3, 6 and 2 are common in Asia (Fig. 1a). Patients with genotype 1 infection have poorer responses to treatment (Pawlotsky, 2003; Zeuzem, 2004) and genotype 1 is common; hence, there is strong interest in a genotype 1 vaccine (Houghton & Abrignani, 2005). However, a vaccine that can provide global protection would be highly desirable. The HIV field faces a similar dilemma, and both single-clade and multi-clade vaccines are being studied in parallel; strategies to cope with diversity include the use of consensus and ancestral sequences, inclusion of only conserved regions, and mosaics (Cattanaro et al., 2006; Fischer et al., 2007; Gaschen et al., 2002; Letourneau et al., 2007; Santra et al., 2008).

Fig. 1(b) compares the genetic variability of HCV and HIV through phylogenetic trees drawn on the same scale. Whilst the overall genetic distances are similar, HIV-1 vaccine efforts only address the HIV-1 M group, not the rare O and N groups. Clearly, a global HIV vaccine needs to cover less diversity than HCV, underscoring the difficulties for the design of a global vaccine for HCV.

In this paper, we will focus first on ways to create an HCV vaccine that optimizes coverage of genotype 1 circulating sequences. We will then address the feasibility and best genomic regions for an HCV global vaccine.

**Designing a genotype 1 vaccine cocktail**

First we designed, analysed and compared the theoretical coverage of different design strategies for a genotype 1 vaccine, considering three general possibilities for single-strain or polyvalent vaccines.

Natural-strain vaccine candidates. We considered a commonly used laboratory genotype 1a strain, H77, as a
prototype and also evaluated other single and polyvalent cocktails of natural strains, both chosen randomly and ones selected computationally to provide the maximal coverage of genotype 1 sequences.

Consensus sequences. Consensus sequences were generated for genotype 1 and subtypes 1a, 1b and 1c, singly and combined, as well as a consensus of subtype consensuses.

Mosaic antigens. Mosaic antigens were designed to maximize coverage of common 9-mers and to minimize the presence of rare epitopes for a given cocktail size. All vaccine strategies employed are listed in Table 1 and the actual sequences are provided in Supplementary Material S1 (available in JGV Online).

**Strategies for genotype 1 vaccine candidates.** We compared 15 distinct candidate vaccine cocktail strategies (Table 1) in terms of two parameters: (i) maximum coverage of 9-mers present in genotype 1 sequences (Fig. 2a, lower graph) and (ii) fewest unique and unnatural 9-mers present (Fig. 2a, upper graph). Of eight single-sequence strategies (H77 as a prototype unselected sequence, the best natural genotype 1a and 1b sequences

![Fig. 1. HCV prevalence in the world and phylogenetic comparison with HIV-1. (a) Geographical distribution of HCV. Shades of brown show the number of infections in different nations based on the population prevalence of HCV (prevalence data: Weekly Epidemiological Record, 8 February 2002, http://www.who.int/docstore/wer/pdf/2002/wer7706.pdf; population sizes: 2006 United Nations Population database, http://www.un.org/esa/population/unpop.htm). Pie charts show the distribution of genotypes in the HCV sequences from nine geographical regions (Los Alamos HCV database). (b) Phylogenetic tree comparison of HCV and HIV genotypes, groups and subtypes. Genetic distances are drawn to the same scale. Trees are based on manually curated representative complete-genome sequence alignments for both viruses. The Findmodel tool was used to determine the best substitution model, which was GTR plus gamma for both datasets. The trees were constructed by using this model in PAUP* (Swofford, 2003).](https://example.com/fig1.png)
in terms of coverage, subtype 1a and 1b consensuses, a consensus of subtype consensuses, a genotype 1 consensus and a mosaic), the reference strain H77 (strategy 1) provided the worst 9-mer coverage (about 50%) of genotype 1 sequences, and the single mosaic sequence (strategy 8) provided the best coverage (about 60%). In addition, natural and consensus sequences had multiple rare 9-mers; the consensus of subtype consensuses was the worst in terms of the number of unnatural 9-mers, whereas the mosaic sequence had no unique or unnatural 9-mers by design. The coverage can be improved significantly if more sequences are used. Cocktails of two to four sequences – natural, consensus or mosaic – provided significant increases in coverage compared with a single sequence. A cocktail of three mosaic sequences (strategy 14) provided 83% coverage; four mosaics were slightly better than three, but the small gain is unlikely to provide discernible benefit or be worth the cost of including the additional antigen.

### Table 1. Vaccine-design strategies

<table>
<thead>
<tr>
<th>Design no.</th>
<th>Description</th>
<th>( k^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotype 1 vaccine designs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>H77 strain</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Natural strain providing best coverage of genotype 1a</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Genotype 1a consensus</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Natural strain providing best coverage of genotype 1b</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Genotype 1b consensus</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Consensus of genotypes 1a, 1b, 1c consensuses</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Genotype 1 consensus</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>One mosaic optimized for genotype 1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>Two natural strains providing best coverage of genotype 1</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>Genotype 1a consensus and genotype 1b consensus</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>Two mosaics optimized for genotype 1</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>Three natural strains providing best coverage of genotype 1</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>Genotype 1a consensus, genotype 1b consensus, genotype 1c consensus</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>Three mosaics optimized for genotype 1</td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>Four natural strains providing best coverage of genotype 1</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>Four mosaics optimized for genotype 1</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>One mosaic optimized for genotype 1a</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>Two mosaics optimized for genotype 1a</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>Three mosaics optimized for genotype 1a</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>Four mosaics optimized for genotype 1a</td>
<td>4</td>
</tr>
<tr>
<td>21</td>
<td>One mosaic optimized for genotype 1b</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>Two mosaics optimized for genotype 1b</td>
<td>2</td>
</tr>
<tr>
<td>23</td>
<td>Three mosaics optimized for genotype 1b</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>Four mosaics optimized for genotype 1b</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>One mosaic optimized for genotypes 1a and 1b</td>
<td>1</td>
</tr>
<tr>
<td>26</td>
<td>Two mosaics optimized for genotypes 1a and 1b</td>
<td>2</td>
</tr>
<tr>
<td>27</td>
<td>Three mosaics optimized for genotypes 1a and 1b</td>
<td>3</td>
</tr>
<tr>
<td>28</td>
<td>Four mosaics optimized for genotypes 1a and 1b</td>
<td>4</td>
</tr>
<tr>
<td>29</td>
<td>Three mosaics serially optimized for genotypes 1b, 1a, 1c</td>
<td>3</td>
</tr>
<tr>
<td>30</td>
<td>Four mosaics serially optimized for genotypes 1b, 1a, 1c, 1</td>
<td>4</td>
</tr>
<tr>
<td>Global vaccine designs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Consensus of genotypes 1, 2, 3, 4, 5, 6 consensuses</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>Five mosaics optimized for global set</td>
<td>5</td>
</tr>
<tr>
<td>33</td>
<td>Consensuses of genotypes 1, 2, 3, 4, 6 combined</td>
<td>5</td>
</tr>
<tr>
<td>34</td>
<td>Five natural strains providing best coverage of genotypes 1, 2, 3, 4, 6</td>
<td>5</td>
</tr>
<tr>
<td>35</td>
<td>Five mosaics serially optimized for genotypes 1, 6, 2, 4, and 3 and 5 combined</td>
<td>5</td>
</tr>
</tbody>
</table>

*\( k^* \), Number of sequences in the cocktail.*
sequences separately (Fig. 2b). Mosaic cocktail vaccines designed for one subtype provided very poor coverage for the other subtype (6–17%), no matter how many sequences were included in the cocktail. This illustrates a probable underlying reason for common cross-subtype reinfection even within the same genotype (Farci & Purcell, 2000; Harcourt et al., 2003; Prince et al., 2005). In contrast, cocktails designed for all of genotype 1 (strategies 8, 11 and 14) showed dramatically better overall coverage, without reducing within-subtype coverage. For example, a three-mosaic cocktail designed for all genotype 1 sequences (strategy 14) gave >80% coverage of 9-mers for both subtypes 1a and 1b (Fig. 2b).

Minimizing sampling frequency bias. As of 375 genotype 1 sequences, only three subtype 1c sequences were available, we adopted the concept of ‘serial optimization’: we optimized one mosaic for subtype 1b, then a second mosaic for subtype 1a, while taking the presence of the 1b-optimized mosaic in the cocktail into account; then a third...
mosaic was optimized for subtype 1c, taking the presence of both other mosaics in the cocktail into account (Fig. 2c, strategy 29). Comparison of trivalent cocktails obtained via regular or serial optimization (Fig. 2c) shows that the coverage of subtypes 1a and 1b is stable with serial optimization, whilst coverage of subtype 1c improves greatly.

Comparison with cocktails of randomly chosen natural sequences. Whilst natural sequences are a more traditional choice for vaccine research than artificial synthetic sequences, reference strains do not provide the best coverage. We therefore compared coverage of genotype 1 provided by one-, two- and three-valent reference, consensus and mosaic cocktails with the distribution of coverage provided by cocktails of the same number of randomly chosen natural sequences (Fig. 2d). Coverage increased with the number of natural sequences in the cocktail, but in all cases, the mosaic sequences provided better coverage than even the best combinations of analogous cocktails of natural or consensus sequences.

Coverage of known CTL epitopes

As our vaccine-optimization strategies are aimed at maximizing vaccine coverage of potential epitopes, an important related question is how well already-known epitopes and their variants are covered. As an example, we show a 24 aa stretch of the E2 protein, MVDPYR-LWHYPCTINTFIFKVRM (Fig. 3). It contains 16 overlapping 9-mers, each of which has multiple variants in the set of 375 genotype 1 sequences. For each overlapping 9-mer, we identify all unique variants present in the reference alignment, calculate the percentage of natural sequences that each unique variant covers and check whether this variant is included in the candidate vaccine cocktail. Comparing reference strain H77 (strategy 1) and the three-mosaic cocktail optimized for genotype 1 (strategy 14), in the first 9-mer array, the variant MVDPYR-LWHYPCTINTFIFKVRM is present in 20% of sequences and is included in both H77 and the three-mosaic cocktail. Two other frequent variants, LVDPYR-LWHYPCTINTFIFKVRM and LVDPYR-LWHYPCINTFYIFKVRM, with combined coverage of 44% of sequences, are not present in H77, but are included in the three-mosaic vaccine cocktail, allowing it to outperform H77 by 44% for this 9-mer. Note that the three overlapping peptides shown are the known HCV CTL epitopes. One of them, HLA-A2-presented epitope RLWHYPCTI (E2 aa 614–622) (Ward et al., 2002) (fourth peptide array from the left, Fig. 3), is present in 66% of the sequences and is included in the three-mosaic cocktail, but is not present in strain H77. The second most frequent variant (27% coverage), which is present both in H77 and in the three-mosaic cocktail, has a V91 mutation in the HLA-A2 anchor residue, thus potentially abrogating HLA-A2 binding. Thus, whilst H77 contains a probable escape variant, the three-mosaic cocktail can potentially prime immune responses to the reactive epitope.

We then assessed the coverage of known HCV CTL epitopes downloaded from the Los Alamos HCV Immunology Database (Yusim et al., 2005; http://www.hcv.lanl.gov) (Supplementary Material S2, available in JGV Online). Fig. 4(a) shows an example of a coverage calculation for NS3 immunodominant epitope KLVALGINAV (Lauer et al., 2004). Of four reported reactive variants (Ward et al., 2002), the H77 strain includes one that is present in 17% of other genotype 1 sequences, whereas the three-mosaic cocktail includes three reactive variants, which together cover 57% of genotype 1 sequences. Supplementary Material S2 contains analogous coverage calculations for other reported CTL epitopes. Overall, even though epitopes tend to be discovered in more conserved regions of the genome (Yusim et al., 2002), a three-mosaic cocktail provided far superior coverage of epitopes and their variants than did strain H77 (P<0.0001, Wilcoxon matched pairs test) (Fig. 4c, d), with improvement in median values of 52% (Fig. 4b). This comparison included only the worst and the best vaccine candidates considered. When the cocktails of three natural strains are compared with three mosaics, however, the mosaics still perform better, as shown in Fig. 2(d), but the benefit varies between 30 and 8% depending on the choice of natural strains (data not shown).

Designing a global HCV vaccine cocktail

The mosaic design methodology can be expanded easily to other genotypes. As a universal vaccine would be preferable to a regional one, we applied the mosaic method to design a vaccine that optimizes coverage for all genotypes. To achieve this, we used an adaptation of the mosaic strategy in which vaccine antigens are added in series to the mosaic cocktail.

Seven strategies were compared for their coverage of the six-genotype target set (Fig. 5): two strategies designed for genotype 1 and five strategies designed to cover the global variability of HCV (Table 1; Supplementary Material S1). The H77 sequence (strategy 1) provides 50% coverage of genotype 1, but only 14–24% coverage of other genotypes. As expected, a trivalent cocktail optimized for genotype 1 (strategy 14) improves the coverage of genotype 1, but not of the other genotypes. A consensus of genotype consensuses (strategy 31) not only provides low coverage of all genotypes, but also has almost 500 unnatural 9-mers, which might result in spurious immune responses (Fig. 5). A five-mosaic cocktail simply optimized for a global set of natural strains (strategy 32) provides very uneven coverage of all genotypes, essentially due to a bias towards genotypes with more sequences (genotypes 1, 2 and 6). To overcome this bias, we included five genotype consensus sequences (for each of the five genotypes with more than two sequences available), five best natural sequences and five mosaics optimized serially for five genotypes, respectively (strategies 33–35; Fig. 5, right of the dashed line). The cocktail of five serial mosaics provided some gain (2–5%) compared with the other two strategies; this cocktail
Fig. 3. Comparison of epitope coverage for two vaccine candidates for genotype 1. As an example of how epitope coverage was calculated, a small portion of alignment is shown for two vaccine candidates: a single natural sequence (strain H77, strategy 1, Table 1) and a three-mosaic cocktail optimized for genotype 1 (strategy 14, Table 1). The alignment in the upper box shows polyprotein positions 608–631 (E2 protein): top line, H77; lower three lines, the three mosaic proteins in the cocktail (strategy 14); dots denote amino acid identity. The lower box shows a subsample of the genotype 1 sequence alignment used as the input. Amino acids that are absent from H77 but present in one or more of the mosaics are shown in red, those present in both are shown in black, and those amino acids not present in either cocktail are shown in blue. Above the alignments are shown some of the overlapping peptide variants present in the natural HCV sequences (potential and known epitopes). In this example, there are three known epitopes (underlined): DYPYRLWHY (HLA-Cw07), RLWHYPCTV (HLA-A2) and TINYTFIK (HLA-A11). For simplicity, only seven of the 16 possible overlapping 9-mers are shown; a single 8-mer array accommodates the 8-mer epitope TINYTFIK. For each overlapping peptide, all variants that are present in >2% of genotype 1 sequences are shown, together with the percentage of sequences in which they are present. As in the alignment sample, peptides absent from H77 but present in one or more of the mosaics are shown in red, those present in neither H77 nor the mosaics are shown in black, and those present in both H77 and the three-mosaic cocktail are shown in blue.
contains no unnatural 9-mers and the fewest unique 9-mers, suggesting that it may be the best candidate for a global vaccine.

### Selecting HCV proteins for use as mosaic antigens

Proteins with relatively low variability have higher intrinsic coverage levels. We examined the relationship between the Shannon entropy of each HCV protein and its coverage by genotype 1-oriented and global vaccines, separately for genotype 1 (Fig. 6a) and genotypes 2–6 (Fig. 6b).

For the conserved HCV Core protein, the five-mosaic global cocktail (strategy 35) provides 91% coverage; as protein variability increases, the coverage provided by the global vaccine drops to 37–38% (Fig. 6a, b). In contrast, the coverage of genotype 1 provided by a three-mosaic vaccine (strategy 14) remains high (75%) for all proteins. This suggests that genotype-specific vaccines might still be effective even when based on more variable proteins.

### Assessing the effect of limited sampling

HCV sequence data for non-1 genotypes are sparse; the sequence variability in the world may be much higher than that observed in the limited set of available complete-genome sequences. To study the effect of more extensive sampling on the apparent coverage of the cocktails, we analysed the Okamoto region, a short (80–100 aa), heavily sequenced region in NS5B. There is almost 10 times more sequencing information for this region than for full-length HCV genomes (Fig. 7b). We compared the complete-genome coverage of our vaccines (strategies 1, 14 and 35) with the coverage of all Okamoto region sequences in the Los Alamos HCV database (Fig. 7a). All vaccine cocktails covered the Okamoto sequences (n=1983 for genotype 1, n=1887 for genotypes 2–6) almost as well as the much smaller complete-genome set, suggesting that, although the vaccine designs were based on the limited set of sequences available, we can expect good coverage of unsampled circulating viruses. As the Okamoto region is quite variable, this comparison provides a conservative estimate

---

**Fig. 4.** Coverage of published CTL epitopes. The coverage of known HCV CTL epitopes and their variants was calculated for strain H77 (strategy 1) and the three-mosaic cocktail (strategy 14). Known CTL epitopes that were defined in humans and shorter than 12 aa were downloaded from the Los Alamos HCV Immunology Database (http://www.hcv.lanl.gov; Yusim et al., 2005). The epitopes were defined by location, meaning that all reactive variants with different presenting HLAs were considered the same epitope if defined within the same protein location, totalling 135 epitopes. For each unique variant of a known epitope, the percentage of 375 genotype 1 sequences in which it is present was calculated. When the variant is included in the vaccine cocktail, it means that the vaccine covers the percentage of sequences that include this variant. (a) An example of epitope coverage calculation for epitope KLVALGINAV (HLA-A2, location NS3 1406–1415; Ward et al., 2002). Published reactive variants are marked by an asterisk. Only variants present in ≥2% of the genotype 1 sequences are shown. H77 coverage, 17.11%; three-mosaic coverage, 56.68%. (b) Epitope coverage with H77 and three mosaics is shown as a scatter plot with median and interquartile range. Each dot represents the coverage of all epitope variants (i.e. all sequences present in the set of 375 genotype 1 sequences) at a given epitope location. The interquartile range is shown by boxes to the right of each scatter plot. (c) Individual epitope coverages are shown by bars (grey, coverage by H77; black, coverage by three-mosaic vaccine cocktail). The epitopes are ordered by their H77 polyprotein sequence position. (d) As (c), except that the epitopes are sorted by the coverage provided with the H77 strain, in descending order.
DISCUSSION

Our first goal was to design a vaccine that would improve the potential epitope coverage of genotype 1. We compared a variety of vaccine reagents, including traditionally used, randomly and rationally selected natural strains, consensus sequences and mosaic antigens. The mosaic vaccine approach was optimal in terms of improving the coverage of both potential epitopes (from 50% using H77 to 83% with a three-antigen mosaic) and reported epitopes (from 39% in median values using H77 to 91% with a three-antigen mosaic). The mosaic design also does not

![Diagram](image)

**Fig. 5.** Comparison of cocktail (and single-strain) vaccine candidates for a globally oriented vaccine. For colours, see the legend to Fig. 2. The lower portion of the graph shows 9-mer coverage of six test sets: 375 genotype 1 sequences, 29 genotype 2 sequences, seven genotype 3 sequences, 11 genotype 4 sequences, three genotype 5 sequences and 42 genotype 6 sequences. The upper portion of the graph shows the number of unnatural and unique 9-mers present in each vaccine candidate. The dashed line separates vaccine cocktails resulting in preferential coverage of genotype 1 and those optimized to improve coverage of all genotypes.

![Diagram](image)

**Fig. 6.** Variability and coverage of HCV proteins. Upper panels (blue), mean Shannon entropy at each position. Lower panels (colours defined in Fig. 2), 9-mer coverage provided by three vaccine candidates. (a) Calculated against 375 genotype 1 sequences; (b) calculated against 92 sequences of genotypes 2–6 combined.

The mosaic vaccine approach was optimal in terms of improving the coverage of both potential epitopes (from 50% using H77 to 83% with a three-antigen mosaic) and reported epitopes (from 39% in median values using H77 to 91% with a three-antigen mosaic). The mosaic design also does not
introduce unnatural or unique epitopes. Next, we investigated the theoretical feasibility of a vaccine that would cover six HCV genotypes. The resulting mosaic designs provided better coverage than consensus or natural-strain vaccines containing the same number of variants, and again contained significantly fewer rare epitopes. Of note, the mosaics’ coverage of 12-mers was very similar to that of 9-mers (not shown), indicating that CD8⁺ T-cell responses are increased relative to natural proteins, so that epitopes that stimulate T-cell responses in the vaccine will be the same epitopes that are processed and presented in natural infection. In vivo experiments are, however, required to validate the approach for HCV.

Experimental studies of mosaics in HIV-1 can be considered as proof of principle: mosaics can be synthesized and expressed and are immunogenic in animal models, and the total number and cross-reactivity of CD8⁺ and CD4⁺ T-cell responses are increased relative to natural and consensus antigens (Barouch et al., 2010; Kong et al., 2009; Santra et al., 2010). The T-cell mosaic antigen insert design could be used together with improvements in other areas of vaccine design. A combination that brings together the best strategies for vector design, adjuvants and antigen design may ultimately enable an effective cross-reactive vaccine-induced immune response against HCV.

**METHODS**

**Hepatitis C sequence data.** To avoid biasing the data with within-patient sequence sets, we used pre-assembled full-length HCV sequence alignments available at the Los Alamos HCV database. These alignments have been verified as having one sequence per patient, and at the time of writing contained 375 genotype 1 sequences (143 subtype 1a, 229 subtype 1b, three subtype 1c), 29 genotype 2, seven genotype 3, 11 genotype 4, three genotype 5 and 42 genotype 6 sequences. Genotype 7 was not included, as only one sequence was available.

**Mosaic vaccine design and epitope coverage.** The algorithms for mosaic vaccine antigen design, as well as the methods that allow selection of natural strains that provide optimal potential epitope coverage, have been described previously (Fischer et al., 2007); a web interface facilitates use of the code (Thurmond et al., 2008). Briefly, the mosaic vaccines are cocktails containing a small number of composite proteins that are designed to optimize representation of potential T-cell epitopes from a set of input viral proteins. Mosaic cocktails are designed from natural sequences in a series of steps that resemble the evolutionary recombination strategies employed by viruses. As the set of experimentally defined HCV epitopes is far from complete, we use fixed-length subsequences as proxies for actual antigens would probably succeed only if based on the most conserved proteins, particularly Core and possibly NS3 and NS4. In contrast, a three-mosaic cocktail optimized for genotype 1 provided reasonable coverage for all HCV genotype 1 proteins, so the variable proteins E1 and E2 can be included in the genotype 1 vaccine, to prime neutralizing antibody responses while providing potentially useful additional T-cell epitopes.

It is important to emphasize that this study addresses maximizing the number of all potential common epitopes, rather than focusing on the important ones; currently it is unclear which HCV-specific T-cell responses confer protection and many HCV epitopes have not yet been mapped. In addition, the intent of this study is to maximize vaccine immunogenicity while circumventing the problems in processing found with polypeptide approaches (reviewed by Korber et al., 2009). Because mosaics are intact proteins that mimic natural proteins, it seems very likely that they will also be processed as natural proteins, so that epitopes that stimulate T-cell responses in the vaccine will be the same epitopes that are processed and presented in natural infection. In vivo experiments are, however, required to validate the approach for HCV.

A serial-optimization strategy to counteract the effects of different sampling frequencies was successful, both in improving the coverage of subtype 1c in the presence of 1a and 1b and in improving the coverage of the global vaccine design. Whilst subtype 1c is relatively rare (Jutavijittum et al., 2009; Verma et al., 2008) and may not need to be included in a vaccine, some genotypes with few available sequences are epidemiologically important. In general, the serial-optimization strategy can be used to improve the coverage of undersampled genotypes without significantly diminishing coverage of the more common genotypes.

Predictably, more conserved proteins have better potential epitope coverage. For a putative global vaccine, the coverage diminishes very rapidly with increased protein variability. A global vaccine cocktail containing five
epitopes; the typical length used is 9 aa, as that is the most common length of CTL epitopes. This optimization criterion achieves near-optimal coverage of epitopes of similar lengths (Korber et al., 2009). The general design strategy is represented schematically in Supplementary Fig. S1 (available in JGV Online) and is as follows. If the goal is to generate a four-valent protein vaccine, four pools of de novo recombinant (mosaic) sequences are created in silico by recombining parental strains chosen randomly from an input sequence set. Recombination breakpoints are chosen so as to preclude the creation of artificial junctional epitopes; that is, in the recombinants, every 9-mer overlapping the recombination breakpoint must be found in the input set of natural sequences, and recombinants that have any unnatural 9-mers are excluded. Meanwhile, the input protein sequence set is fractured into all possible 9-mers, and the tally of each 9-mer is calculated to provide a basis for evaluating potential epitope coverage of vaccine candidates. The program then tests all combinations of four mosaics, one drawn from each pool, and selects the set of four that maximizes the coverage of all 9-mers in the input set of sequences. Finally, new mosaics are introduced into each sample pool, and the recombination and optimization processes are repeated iteratively until the 9-mer coverage of the polyvalent mosaic vaccine set no longer improves (the run time is typically 1–2 days).

For coverage assessment to compare vaccine cocktail candidates, we calculate the number of perfect 9-mer matches between the vaccine cocktail and a population of interest. The overall vaccine coverage is computed as a mean of the coverage for every sequence in the input dataset; i.e. the proportion of covered 9-mers is computed for each natural sequence, and the overall coverage is the mean of these values. Additionally, because slightly mismatched epitopes often retain some natural sequence, and the overall coverage is the mean of these values. Therefore, to calculate partial 8/9 and 7/9 matches, we also calculate partial 8/9 and 7/9 matches.

Measure of variability. Shannon entropy is a diversity measure that incorporates both the number of possible amino acids allowed and their frequency as:

\[- \sum_{aa} P(aa) \log_{10}[P(aa)]\]

where \(P(aa)\) is the proportion of each amino acid in column \(i\) of a sequence alignment (Korber et al., 1994; Yusim et al., 2002). Shannon entropy was calculated for each position in the protein alignment as a representative measure of diversity for Fig. 6.

ACKNOWLEDGEMENTS

Funding for this work was provided by the Los Alamos National Laboratory Directed Research Program (LDRD 20050155DR, ‘Rational Vaccine Design’). We thank Jennifer Macke for assistance in editing the manuscript.

REFERENCES


Addressing HCV diversity for an optimal vaccine


