RXLR effector genes (15, 23). Moreover, Hpa effector genes generally were not located in syntenic locations relative to Phytophthora genomes, except for three families of effectors, which have unusually high levels of sequence conservation (Fig. 2).

As obligate biotrophs, downy mildews may have lost some metabolic pathways. We identified several potential metabolic defects in Hpa compared with P. sojae and P. ramorum (fig. S9). For example, genes for nitrate and nitrite reductases, a nitrate transporter, and sulfite reductase were missing (fig. S10 and table S3), which is also a feature of the genomes of obligately parasitic powdery mildew fungi (24). Hpa also lacks genes required for synthesis of arachidonic acid and polyamine oxidases.

Flagellated zoospores are produced by many oomycetes (25). Contrastingly, several downy mildew lineages germinate by extending infective germ tubes from nonmotile conidiospores, although evidence exists for a rare zoospore stage in some other conidial downy mildews (26, 27). To conclusively determine whether spore motility has been lost from the Hpa lineage, we searched the Hpa genome for 90 flagella-associated genes using Chlamydomonas sequences and their Phytophthora orthologs (28). No matches were detected in Hpa for any of these. Similarly, many Phytophthora adhesion-related genes are reduced in number or absent from Hpa, consistent with the lack of adherent cysts that normally develop from zoospores during infection.

Analysis of Hpa gene space revealed genomic signatures of major alterations in pathogenic strategy, metabolism, and development that occurred during the evolution of obligate biotrophy from a facultative, hemibiotrophic ancestor. Interestingly, some features of Hpa gene space (large numbers of secreted effectors, reduction in degradative enzymes, and loss of N and S assimilation) are mirrored in genomes of biotrophic fungi (24, 29, 30). These similarities indicate that convergent adaptations occurred during the independent evolution of biotrophy in fungal and oomycete lineages.

References and Notes

10. Materials and methods are available as supporting material on Science Online.
18. S. Kale et al., Cell 142, 284 (2010).
31. We thank E. Holub for providing the Emoy2 isolate, D. Greenshields and N. Bruce for technical assistance, A. Heck and M. Slizier for analysis of secreted Hpa proteins, R. Hubley for creating repeat modeller libraries, and participants in the 2007 Annotation Jamboree and in the 2008 and 2009 Oomycete Bioinformatics Training Workshops for sequence annotations. This research was supported by grants EF-0412213, I0S-0744875, I0S-0924861, and MCB-0639266 from the U.S. NSF and 2004-35600-15055 and 2007-35319-18100 from the U.S. Department of Agriculture National Institute of Food and Agriculture to B.M.T. and J.M.M.; Biotechnology and Biological Sciences Research Council BBSRC BB/C59213/1, BB/E0248821, and Engineering and Physical Sciences Research Council/BBSRC Systems Biology DTC student EF/F050002/S1 to J.B.; Gatsby GAT2545 and BBSRC BB/F0161901, BB/E0248821, and BBSRC CASE studentship T12144 to J.D.G.J. Other support is detailed in the supporting online material. Genome browsers are maintained at the Virginia Bioinformatics Institute (vmb.vbi.vt.edu) and the Sainsbury Laboratories (gbrowse2.tsl.ac.uk/cgi-bin/gb2/gbrowse/hpa_emo2_publication).

Supporting Online Material

www.sciencemag.org/cgi/content/full/330/6010/1549/DC1
Materials and Methods
Figs. S1 to S10
Tables S1 to S3
References
16 July 2010; accepted 25 October 2010
10.1126/science.1195203

The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation

The International HIV Controllers Study†‡

Infectious and inflammatory diseases have repeatedly shown strong genetic associations within the major histocompatibility complex (MHC); however, the basis for these associations remains elusive. To define host genetic effects on the outcome of a chronic viral infection, we performed genome-wide association analysis in a multiethnic cohort of HIV-1 controllers and progressors, and we analyzed the effects of individual amino acids within the classical human leukocyte antigen (HLA) proteins. We identified >300 genome-wide significant single-nucleotide polymorphisms (SNPs) within the MHC and none elsewhere. Specific amino acids in the HLA-B peptide binding groove, as well as an independent HLA-C effect, explain the SNP associations and reconcile both protective and risk HLA alleles. These results implicate the nature of the HLA–viral peptide interaction as the major factor modulating durable control of HIV infection.

HIV infection is characterized by acute viremia, often in excess of 5 million viral particles per milliliter of plasma, followed by an average 100-fold or greater decline to a relatively stable plasma virus load set point (I). In the absence of antiretroviral therapy, the level of viremia is associated with the rate of CD4+ T cell decline and progression to AIDS. There is substantial interperson variability in the virus load set point, with most individuals having stable levels exceeding 10,000 RNA copies/ml. Yet a small number of people demonstrate sustained ability to control HIV replication without therapy. Such individuals, referred to as HIV controllers, typically maintain stable CD4+ cell counts, do not develop clinical disease, and are less likely to transmit HIV to others (2).

To determine the genetic basis for this rare phenomenon, we established a multinational consortium (www.hivcontrollers.org) to recruit HIV-1 controllers, who are defined by at least three measurements of plasma virus load (VL) < 2000 RNA copies/ml over at least a 12-month period in the absence of antiviral therapy. We performed a genome-wide association study (GWAS) in the HIV controllers (median VL, CD4 count, and disease duration of 241 copies/ml, 699 cells/mm3, and 10 years, respectively) and treatment-naïve chronically infected individuals with advanced disease (median VL and CD4 count of 61,698 copies/ml and 224 cells/mm3, respectively) enrolled in antiviral treatment studies led by the AIDS Clinical Trials Group. After quality control and imputation on the basis of HapMap

"All authors with their contributions and affiliations appear at the end of this paper.
†To whom correspondence should be addressed. E-mail: bwalker@partner.org (B.D.W.); pdebakker@rics.bwh.harvard.edu (P.I.W.d.B.)

www.sciencemag.org SCIENCE VOL 330 10 DECEMBER 2010
Phase 3 (3), we obtained data on 1,384,048 single-nucleotide polymorphisms (SNPs) in 974 controllers (cases) and 2648 progressors (controls) from multiple populations (table S1).

After stratification into European, African American, and Hispanic ethnic groups (fig. S1), we tested each SNP for association using logistic regression, including the major principal components as covariates to correct for population substructure (4). In the largest group, comprising 1712 individuals of European ancestry, we identified 313 SNPs with genome-wide significance, defined by \( P < 5 \times 10^{-8} \) due to correction for multiple comparisons (table S2). All SNPs that reached genome-wide significance were located in the major histocompatibility complex (MHC) region on chromosome 6 (Fig. 1A). We obtained similar results for the other two ethnic groups and in a meta-analysis of all participants (fig. S2). We also performed a genome-wide analysis to test the influence of local chromosomal ancestry in the African American sample (4), but we detected no signal outside the MHC (figs. S3 and S4). The impact of the MHC was further underscored when we specifically tested published associations related to HIV disease progression outside the MHC. Only variants in the \( CCR5-CCR2 \) locus—namely, \( CCR5A32 \) deletion polymorphism (5), \( 927T \) in \( CCR5 \) (6), and \( Val^{36} \to Ile^{34} \) in \( CCR2 \) (7)—replicate with nominal statistical significance in our study (Fig. 1B and table S3).

Closer examination of the significant SNPs within the MHC showed that they were located in a 3-Mb region concentrated around class I human leukocyte antigen (HLA) genes (fig. S5), but extensive linkage disequilibrium (LD) makes precise assignment of causal variants challenging (8). Therefore, we used stepwise regression to define independent markers associated with host control. From the initial set of 313 SNPs that reached genome-wide significance in the European sample, for which the greatest number of participants were available, we found only four independent markers of association (Table 1). rs9264942, located 35 kb upstream of \( HLA-C \) and a putative variant associated with \( HLA-C \) expression levels [odds ratio (OR) = 2.9, \( P = 2.8 \times 10^{-35} \), where an OR > 1 indicates a protective effect], and rs2395029, a proxy for \( HLA-B^{*}57:01 \) (OR = 5.3, \( P = 9.7 \times 10^{-26} \)), had been previously reported to be associated with virus load set point after acute infection (9). We also defined rs4418214, a noncoding SNP near \( MICA \) (OR = 4.4, \( P = 1.4 \times 10^{-34} \)), and rs3131018 in \( PSORS1C3 \), a gene implicated in psoriasis (OR = 2.1, \( P = 4.2 \times 10^{-16} \)). These four SNPs explain 19% of the observed variance of host control in the European sample; together with those in \( CCR5 \), these SNPs explain 23%, using Nagelkerke’s approximation (Fig. 1C) (10).

In the smaller African American sample, we observed 33 SNPs with genome-wide significance, four of which were identified as independent markers, but all differed from those in the European sample (Table 1). This suggests that shared causal variants are tagged by different SNPs in these two populations or that the mechanism of control differs with ethnicity. Only rs2523608 was previously identified, in a recent study of virus load set point in African Americans (11). Despite no evidence for historical recombi-

![Fig. 1. Genome-wide association results in the European sample. (A) Manhattan plot of 1.3 million autosomal SNPs. Only SNPs in the MHC on chromosome 6 reach genome-wide significance, indicated by the horizontal dotted line (\( P < 5 \times 10^{-8} \)). Red and blue colors alternate between chromosomes. (B) Quantile-quantile plot of the association results with (black) and without (blue) SNPs in the extended MHC and the \( CCR5-CCR2 \) locus, indicating that the detectable effect is entirely attributable to these two loci. The red line denotes the expected distribution under the null hypothesis of no effect. (C) Distribution of the genotype protective score, defined as the total number of alleles associated with host control at the four independent SNPs in the MHC and the variants at \( CCR5-CCR2 \), showing marked differences in controllers (orange) and progressors (blue). In aggregate, these variants explain 23% of the observed variance of durable host control.](https://www.sciencemag.org)

<table>
<thead>
<tr>
<th>SNP</th>
<th>A1</th>
<th>A2</th>
<th>Frequency in controllers</th>
<th>Frequency in progressors</th>
<th>Univariate OR</th>
<th>Univariate P value</th>
<th>Multivariate OR</th>
<th>Multivariate P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs9264942</td>
<td>C</td>
<td>T</td>
<td>0.595</td>
<td>0.336</td>
<td>2.9</td>
<td>2.8 × 10^{-35}</td>
<td>2.1</td>
<td>6.3 × 10^{-16}</td>
</tr>
<tr>
<td>rs4418214</td>
<td>C</td>
<td>T</td>
<td>0.240</td>
<td>0.075</td>
<td>4.4</td>
<td>1.4 × 10^{-34}</td>
<td>1.8</td>
<td>4.9 × 10^{-4}</td>
</tr>
<tr>
<td>rs2395029</td>
<td>G</td>
<td>T</td>
<td>0.139</td>
<td>0.032</td>
<td>5.3</td>
<td>9.7 × 10^{-26}</td>
<td>2.1</td>
<td>3.5 × 10^{-5}</td>
</tr>
<tr>
<td>rs3131018</td>
<td>C</td>
<td>A</td>
<td>0.777</td>
<td>0.625</td>
<td>2.1</td>
<td>4.2 × 10^{-16}</td>
<td>1.5</td>
<td>1.2 × 10^{-5}</td>
</tr>
<tr>
<td>rs2523608</td>
<td>G</td>
<td>A</td>
<td>0.522</td>
<td>0.326</td>
<td>2.6</td>
<td>8.9 × 10^{-20}</td>
<td>2.3</td>
<td>3.7 × 10^{-15}</td>
</tr>
<tr>
<td>rs2523590</td>
<td>T</td>
<td>G</td>
<td>0.264</td>
<td>0.137</td>
<td>2.7</td>
<td>3.5 × 10^{-14}</td>
<td>1.9</td>
<td>2.1 × 10^{-6}</td>
</tr>
<tr>
<td>rs3131018</td>
<td>C</td>
<td>T</td>
<td>0.300</td>
<td>0.164</td>
<td>2.4</td>
<td>1.7 × 10^{-13}</td>
<td>2.3</td>
<td>1.2 × 10^{-12}</td>
</tr>
<tr>
<td>rs9262632</td>
<td>G</td>
<td>A</td>
<td>0.097</td>
<td>0.034</td>
<td>3.1</td>
<td>1.0 × 10^{-8}</td>
<td>2.2</td>
<td>2.8 × 10^{-4}</td>
</tr>
</tbody>
</table>

Table 1. Association results for the independent SNPs in the MHC identified with stepwise regression in the European and African American samples. The odds ratio and frequency is given for the A1 allele, where OR > 1 indicates a protective effect. Odds ratios and P values were computed for univariate and multivariate regression models.
tion ($D^2 = 1$), this SNP is only weakly correlated ($r^2 < 0.1$) with HLA-B*57:03, the class I allele most strongly associated with durable control of HIV in populations of African ancestry (11–13).

In the Hispanic sample, which was much smaller, the most significant SNP was rs2523590, 2 kb upstream of HLA-B, also identified in the African American sample described here.

Given the localization of significant SNPs entirely to the HLA class I region, as well as previous studies showing HLA alleles to affect disease progression (13–20), we next sought to evaluate whether these SNP and HLA associations might be due to specific amino acids within HLA. Because HLA types were available for only a portion of the entire cohort, we developed a method to impute classical HLA alleles and their corresponding amino acid sequences (4) on the basis of haplotype patterns in an independent data set collected by the Type 1 Diabetes Genetics Consortium (T1DGC) (21). This data set contains genotype data for 639 SNPs in the MHC that overlap with genotyped SNPs in our GWAS and classical HLA types for class I and II loci at four-digit resolution in 2767 unrelated individuals of European descent.

We imputed HLA types in the European sample of our study and validated the imputations by comparing to empirical four-digit HLA typing data collected for class I loci in a subset (n = 371) of the HIV controllers. The quality of the imputations was such that the imputed and true frequencies for all HLA alleles in this subset were in near-perfect agreement (Fig. 2A) ($r^2 = 0.99$). Furthermore, the positive predictive value was 95.2% and the sensitivity was 95.2% at two-digit resolution (92.7 and 95.6%, respectively, at four-digit resolution) for HLA alleles with frequency >2% (Fig. 2B). This indicates that the performance of the imputation was generally excellent for common alleles, consistent with previous work (22). We used HLA allele imputations in all participants (even those with HLA types defined by sequencing) for association analyses to avoid systematic bias between cases and controls. Lower imputation quality would only decrease power, not increase the false-positive rate, because cases and controls would be equally affected.

We tested all HLA alleles for association via logistic regression, adjusting for the same covariates used in SNP analysis (tables S4 and S5). The most significant HLA association is B*57:01 (OR = 5.5, $P = 1.4 \times 10^{-28}$), which explains the proxy association of rs2395029 in HCP5. With the use of stepwise regression modeling in the European sample of controllers and progressors, we were able to implicate B*57:01, B*27:05, B*14/Cw*08:02, B*52, and A*25 as protective alleles and B*35 and Cw*07 as risk alleles. These associations are consistent with earlier studies that highlighted a role for HLA class I alleles (13–20), and particularly HLA-B alleles in control of HIV, which indicated that the imputations are robust. Collectively explaining 19% of the variance of host control, these HLA allele associations are consistent with the effects of the four independent SNPs.

Virus-infected cells are recognized by CD8+ T cells after presentation of short viral peptides within the binding groove of HLA class I, and HIV-specific CD8+ T cells are strongly associated with control (23). We thus evaluated whether the SNP associations identified in the GWAS, and the HLA associations derived from imputation, might be due to specific amino acid positions within the HLA molecules, particularly those involved in the interaction between the viral peptide and the HLA class I molecule. Using the official DNA sequences defined for known HLA alleles (24), we encoded all variable amino acid positions within the coding regions of the HLA genes in each of the previously HLA-typed 2767 individuals in the T1DGC reference panel, and we used this data set to impute the amino acids in the cases and controls (4). Among a total of 372 polymorphic amino acid positions in class I and II HLA proteins, 286 are biallelic like a typical nonsynonymous coding SNP. The remaining 86 positions accommodate more than two amino acids; position 97 is the most diverse in HLA-B with six possible amino acids observed in European populations.

After imputing these amino acids in the European sample, we used logistic regression to test all positions for association with host control (fig. S6 and table S6). Notably, position 97 in HLA-B was more significant (omnibus $P = 4 \times 10^{-45}$) than any single SNP in the GWAS, and three amino acid positions (67, 70, and 97), all in HLA-B, showed much stronger associations than any single classical HLA allele, including B*57:01 (Fig. 3A). Moreover, allelic variants at these positions were associated with substantial frequency differences between cases and controls (Fig. 3B). These results indicate that the effect of HLA-B on disease outcome could be mediated, at least in part, by these positions. These three amino acid positions are located in the peptide binding groove, which suggests that conformational differences in peptide presentation at these sites contribute to the protective or susceptible nature of the various HLA-B allotypes. Although both innate and adaptive mechanisms could be at play, the hypothesis that HLA affects peptide presentation and subsequent T cell functionality is supported by experimental data showing substantial functional differences between CTL targeting identical epitopes but restricted by different HLA alleles (25).

We next performed stepwise regression modeling and identified six residues as independent markers associated with durable control of HIV. These include Arg67, Cys67, Gly62, and Glu63, all in HLA-B; Ser27 in HLA-A; and Met104 in HLA-C, which collectively explain 20% of the observed variance (similar to the variance explained by the seven classical HLA alleles described above). With the exception of Met104 in the transmembrane domain of HLA-C, these residues are all located in the MHC class I peptide binding groove, again suggesting that the binding pocket—and, by inference, the conformational presentation of class I-restricted epitopes—plays a key role in host control.

Having identified these amino acid positions as strong candidates to account for the SNP and HLA association signals in this study, we next investigated their effects on protection or risk, revealing allelic variants at these positions linked to both extremes (Table 2). HLA-B position 97 (omnibus $P = 4 \times 10^{-45}$), located at the base of the C pocket, has important conformational properties for peptide binding (26). Position 97 has six allelic variants: Protective haplotypes B*57:01, B*27:05, and B*14 are uniquely defined by Val87.
risk haplotypes Cw*07, B*07, and others, whereas Thr97 (11%) lies on protective B*52 (and others). Arg97 is the most common amino acid (51%) and is carried by risk allele B*35, among others.

The importance of this amino acid position to host control is underscored by conditional analyses revealing significance when we adjust incrementally for Val97 (omnibus test for position

**Fig. 3.** Associations at amino acids in HLA-B in the European sample. (A) Association results for all variable amino acid positions, as calculated by the omnibus test. Colors denote conventional pocket positions. P values for significant classical HLA-B alleles are shown for comparison. (B) Marked allele frequency differences between controllers and progressors for amino acids at positions 67, 70, and 97. Numbers above the bars indicate odds ratios (values >1 indicate a protective effect). (C) Associations between allelic variants at amino acid positions 67, 70, and 97 and quantitative virus load set point in the independent Swiss HIV cohort study. Effect estimates (beta coefficients from a linear-regression model) are given in log_{10} units of virus load set point. P values refer to the omnibus test for association at each position. Error bars indicate the standard error of the beta coefficient.

### Table 2. Haplotypes defined by the four independent SNPs, classical HLA alleles, and amino acids associated with host control in the European sample. Haplotypes are ordered by the estimated odds ratio, where the most common haplotype was taken as reference (OR = 1). P values are for each haplotype tested against all other haplotypes. Only haplotypes with >1% frequency are listed, accounting for >85% of haplotype diversity. HLA-A alleles were excluded to limit the number of haplotypes. See (33).

<table>
<thead>
<tr>
<th>rs3131018</th>
<th>HLA-C</th>
<th>rs9264942</th>
<th>HLA-B</th>
<th>rs4418214</th>
<th>rs2395029</th>
<th>Frequency</th>
<th>OR</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>M</td>
<td>C</td>
<td>B*57:01</td>
<td>G</td>
<td>E</td>
<td>M</td>
<td>S</td>
<td>V</td>
</tr>
<tr>
<td>C</td>
<td>M</td>
<td>C</td>
<td>B*52:01</td>
<td>R</td>
<td>E</td>
<td>S</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>C</td>
<td>V</td>
<td>C</td>
<td>B*27:05</td>
<td>R</td>
<td>E</td>
<td>C</td>
<td>K</td>
<td>N</td>
</tr>
<tr>
<td>C</td>
<td>M</td>
<td>C</td>
<td></td>
<td>R</td>
<td>E</td>
<td>S</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>C</td>
<td>Cw*08:02</td>
<td>M</td>
<td>C</td>
<td>B*14:02</td>
<td>R</td>
<td>N</td>
<td>C</td>
<td>N</td>
</tr>
<tr>
<td>C</td>
<td>M</td>
<td>C</td>
<td></td>
<td>R</td>
<td>N</td>
<td>S</td>
<td>N</td>
<td>R</td>
</tr>
<tr>
<td>C</td>
<td>V</td>
<td>C</td>
<td></td>
<td>R</td>
<td>N</td>
<td>F</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>C</td>
<td>M</td>
<td>C</td>
<td></td>
<td>R</td>
<td>N</td>
<td>C</td>
<td>N</td>
<td>R</td>
</tr>
<tr>
<td>C</td>
<td>V</td>
<td>C</td>
<td></td>
<td>R</td>
<td>E</td>
<td>S</td>
<td>N</td>
<td>S</td>
</tr>
<tr>
<td>A</td>
<td>M</td>
<td>C</td>
<td></td>
<td>R</td>
<td>E</td>
<td>S</td>
<td>N</td>
<td>R</td>
</tr>
<tr>
<td>C</td>
<td>M</td>
<td>T</td>
<td></td>
<td>R</td>
<td>N</td>
<td>F</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>A</td>
<td>M</td>
<td>C</td>
<td></td>
<td>R</td>
<td>N</td>
<td>S</td>
<td>N</td>
<td>R</td>
</tr>
<tr>
<td>C</td>
<td>V</td>
<td>T</td>
<td></td>
<td>R</td>
<td>E</td>
<td>S</td>
<td>N</td>
<td>R</td>
</tr>
<tr>
<td>C</td>
<td>Cw*07:01</td>
<td>M</td>
<td>C</td>
<td></td>
<td>R</td>
<td>N</td>
<td>S</td>
<td>N</td>
</tr>
<tr>
<td>A</td>
<td>V</td>
<td>T</td>
<td></td>
<td>R</td>
<td>E</td>
<td>S</td>
<td>N</td>
<td>R</td>
</tr>
<tr>
<td>C</td>
<td>Cw*07:01</td>
<td>V</td>
<td>T</td>
<td>B*08:01</td>
<td>R</td>
<td>N</td>
<td>F</td>
<td>N</td>
</tr>
<tr>
<td>A</td>
<td>V</td>
<td>T</td>
<td></td>
<td>R</td>
<td>N</td>
<td>Y</td>
<td>Q</td>
<td>T</td>
</tr>
<tr>
<td>A</td>
<td>Cw*07:02</td>
<td>V</td>
<td>T</td>
<td>B*07:02</td>
<td>R</td>
<td>N</td>
<td>Y</td>
<td>Q</td>
</tr>
<tr>
<td>A</td>
<td>V</td>
<td>T</td>
<td></td>
<td>R</td>
<td>N</td>
<td>F</td>
<td>N</td>
<td>R</td>
</tr>
<tr>
<td>A</td>
<td>V</td>
<td>T</td>
<td></td>
<td>R</td>
<td>N</td>
<td>F</td>
<td>N</td>
<td>R</td>
</tr>
</tbody>
</table>
97, P = 3 × 10−20), Asn97 (P = 2 × 10−9), and Trp92 (P = 7 × 10−5). Thus, at a single position within the peptide binding groove (position 97, C-pocket), discrete amino acids are associated with opposite disease outcomes, even after controlling for B*57 and B*27, alleles associated with host control.

We also found similar discordant associations for alleles at positions 67, 63, and 62 (Table 2), all of which line the α1 helix along the peptide binding groove and help shape the B-pocket (Fig. 4). At position 67 (omnibus P = 2 × 10−5), risk haplotypes B*57 and B*07 carry aromatic residues Phe67 and Tyr67, respectively, whereas protective B*57:01, B*27:05, and B*14 alleles carry sulfur-containing residues Met67 or Cys67. Position 62 (P = 5 × 10−7) is biallelic (Arg/Gly) with the Gly62 allele segregating with protective alleles B*57:01 and B*58 (<1% frequency, OR = 1.7, P = 0.2). Adjacent position 63 (P = 9 × 10−16) is also biallelic (Glu/Asn) with Glu63 appearing in complete LD (|D′| = 1) with B*57:01, B*27:05, and B*52. In contrast, at this position the risk alleles B*07 (14% frequency, OR = 0.5, P = 1 × 10−5) and B*35 both carry Asn63. Position 70 (omnibus P = 3 × 10−39) accommodates four alleles that are tightly coupled with positions 67 and 97: Ser70 appears exclusively with Met67 (which defines B*57 and B*58), Glu70 with Tyr67, and Lys70 with Asn97 (B*27). Hence, these data create a consistent and parsimonious model that can explain the associations of classical HLA-B alleles by specific amino acids lining the binding groove (and residues tightly coupled to them), which are expected to have an impact on the three-dimensional structure of the peptide-MHC complex.

To further investigate the role of individual amino acid positions in HLA-B, we implemented a permutation procedure to assess how consistent the above observations are with a null model in which there is no relation between amino acids at a particular position and host control (4). The results of this procedure provided evidence that multiple amino acid positions in the peptide binding groove are indeed associated with host control (table S7), including positions 62, 63, 67, 70, and 97, thus providing a structural basis for the effect of HLA-B on host control (Fig. 4).

Within HLA-A position 77, which lies on the α helix contributing to the F-pocket, we identified a weaker but still significant association (omnibus P = 3 × 10−6). Ser77 (6% frequency, OR = 2.0, P = 2 × 10−5) is carried by only two HLA-A alleles (joint r2 = 1). A*44 (2.4% frequency, OR = 2.6, P = 1 × 10−5) and A*32 (3.2%, OR = 1.6, P = 0.02). Given its location and earlier association evidence for the A10 supertype (27), HLA-A could play a role in host control, although the evidence is not as strong as for HLA-B.

The signals within HLA-C are less straightforward to interpret. Position 304 is a biallelic variant (Val/Met) located in the transmembrane domain (Met304, 28% frequency, OR = 2.3, P = 7 × 10−42), Asn97 (6% frequency, P = 2.8 × 10−4). Similarly, addition of rs9264942 to a multivariate model of all seven independent classical HLA alleles, and amino acids in a second independent cohort of untreated HIV-infected persons from Switzerland (fig. S7 and tables S8 and S9) (4), in which virus load set point was measured as a quantitative trait. Allelic variants at positions 67, 70, and 97 were also associated with highly significant differences in virus load set point in this second cohort (Fig. 3C). The effect estimates of all variable amino acids defined here are consistent with the major genetic effects are condensed to single positions (omnibus P = 1 × 10−11). The HLA-A associations (A*44 or Ser77) did not replicate, which reduces the likelihood that HLA-A plays a major role in host control.

In the African American sample (fig. S10), the most significant HLA-allele association was observed for two-digit B*57 (OR = 5.1, P = 1.7 × 10−21) and four-digit B*57:03 (OR = 5.1, P = 2.8 × 10−17; tables S10 and S11), consistent with previous studies (11–13). Position 97 in HLA-B (omnibus P = 2 × 10−25) is again the most significant amino acid (table S12). The consistency of these results demonstrates that imputation and association testing at amino acid resolution in multiple ethnicities can resolve disparate SNP associations in the MHC and help with fine-mapping of classical HLA associations.

Altogether, these results link the major genetic impact of host control of HIV-1 to specific amino acids involved in the presentation of viral peptides on infected cells. Moreover, they reconcile previously reported SNP and HLA associations with host control and lack of control to specific amino acid positions within the MHC class I peptide binding groove. Although variation in the entire HLA protein is involved in the differential response to HIV across HLA allotypes, the major genetic effects are conditioned to the positions highlighted in this study, indicating a structural basis for the HLA association with disease progression that is probably mediated by the conformation of the peptide within the class I binding groove. The most significant residue, position 97 in the floor of the peptide binding groove of HLA-B, is associated with the extremes of viral load, depending on the expressed amino acid. This residue has been shown to have important conformational properties that affect epitope-contacting residues within the binding groove (26, 30) and has also been implicated in HLA protein folding and cell-surface expression (31).

Although the main focus of this study was on common sequence variation, it remains an open question as to the role of variants outside the MHC and the contribution of epistatic effects and epigenetic regulation. Additional factors also contribute to immune control of HIV, including fitness-altering mutations, immunoregulatory networks, T cell help, thymic selection, and innate effector mechanisms such as killer cell immunoglobulin-like receptor recognition (23), some of which are influenced by the peptide-HLA class I complex. However, the combination and location of the significant amino acids defined here are most consistent with the genetic associations observed being modulated by HLA class I restricted CD8+ T cells. These results implicate the nature of the HLA–viral peptide interaction as the major genetic factor modulating durable control of HIV infection and provide the basis for future studies of the impact of HLA-peptide conformation on immune cell induction and function.

References and Notes


4. See supporting online material on Science Online for detailed background on the analyses that we performed.


Fig. 4. Three-dimensional ribbon representation of the HLA-B protein based on Protein Data Bank entry 2bvp (30), highlighting amino acid positions 62, 63, 67, 70, and 97 lining the peptide binding pocket. The peptide backbone of the epitope is also displayed. This figure was prepared with UCSF Chimera (32).