Signatures of strong population differentiation shape extended haplotypes across the human *CD28*, *CTLA4*, and *ICOS* costimulatory genes

Vincent Butty*, Matt Roy*, Pardis Sabeti†, Whitney Besse*, Christophe Benoist*‡, and Diane Mathis*‡

*Section on Immunology and Immunogenetics, Joslin Diabetes Center; Department of Medicine, Brigham and Women's Hospital; Harvard Medical School, Boston, MA 02215; and †The Broad Institute, Harvard University and Massachusetts Institute of Technology, Cambridge, MA 02142

Contributed by Christophe Benoist, November 16, 2006 (sent for review October 27, 2006)

The three members of the costimulatory receptor family, CD28, CTLA-4, and ICOS, have complementary effects on T cell activation, and their balance controls the overall outcome of immune and autoimmune responses. They are encoded in a short genomic interval, and overall activity may result from interplay between allelic variants at each locus. With multiethnic DNA panels that represent a wide spectrum of human populations, we demonstrate long-range linkage disequilibrium among the three genes. A large fraction of the variation found in the locus can be explained by the presence of extended haplotypes encompassing variants at *CD28*, *CTLA4*, and the *ICOS* promoter. There are unusual differences in the distribution of some variants and haplotypes between geographic regions. The differences may reflect demographic events and/or the adaptation to diverse environmental and microbial challenges encountered in the course of human migrations and will be important to consider when interpreting association to immune/autoimmune responsiveness.

autoimmunity | linkage disequilibrium | T cell costimulation

Proper balance of immune responses is key to guarantee adequate host protection with minimal inflammatory and immunopathological damage. The activation of naïve T lymphocytes requires antigen-specific signals, complemented by non-antigen-specific "costimulatory" signals that modulate the extent of activation and its phenotypic outcome. Members of the *CD28* gene family are major costimulators, through their interactions with molecules of the *B7* family on antigen-presenting and stromal cells (reviewed in ref. 1). Three members of the *CD28* family, CD28, CTLA-4, and ICOS, have overlapping and complementary functions. CD28–B7 interactions contribute a positive signal toward a productive activation of T cells and amplify the immune response; in contrast, CTLA-4 acts as a counterregulatory molecule, dampening the response through a variety of means. The action of ICOS is subtler: it has some positive costimulatory ability, and ICOS-positive activated cells have potent effector activity, yet *ICOS* is also strongly expressed and active on T regulatory cells. ICOS regulates cytokine secretion patterns (Th1/Th2 balance) during infections and in settings of autoimmunity or atopy (1).

The *CD28*, *CTLA4*, and *ICOS* genes lie within a stretch of 300 kb on human chromosome 2, a configuration most probably resulting from sequential duplications. Their expression is differentially regulated: although CD28 is constitutively present on naïve T cells, CTLA-4 and ICOS are displayed only after activation, through transcriptional induction for both and/or intracellular redistribution (2). In keeping with their intertwined function, their expression is also interlocked, with CD28 engagement influencing the expression of the other two molecules (2).

Natural genetic variation in the region seems to exert an impact on autoimmune diseases such as type 1 diabetes, Graves' disease, or multiple sclerosis in human patients and in the corresponding animal models (3). There is a high degree of conservation of the proteins, with essentially no variability in the coding regions within species, suggesting that it must be variation in gene transcription, splicing, or transcript stability that is associated with autoimmune susceptibility. Recent murine data suggest that extended costimulatory haplotypes with differential regulation of both *CTLA4* and *ICOS* partition with autoimmune susceptibility (4).

More than 80% of the human genome is organized in "haplotype blocks" of high linkage disequilibrium (LD), resulting from a combination of population genetics events (e.g., bottlenecks) and punctuated variation in the recombination rate (cold spots and hot spots) (5). These blocks of high local LD cover a few to a few tens of kilobases and are shorter in African populations, consistent with a longer evolutionary history and greater accumulation of recombination events. The blocks encompass a limited number of major variants (typically 4–7).

Our goals in this study were to investigate the structure of combined variation in the human costimulatory locus in a worldwide panel and to seek signatures of adaptation that could have an impact on the susceptibility to autoimmune diseases. In particular, given the complementary and balancing functions of the *CD28* family members and their cross-regulation, we hypothesized that functional variation in the *CD28*/*CTLA4*/*ICOS* region may not be due to polymorphism in a single gene but to the combination of genetic variants at the three loci.

Results

To obtain a broad view of the costimulatory locus in diverse human populations we used the multiethnic sample collection of the Human Genome Diversity Panel (HGDP) (6, 7), which includes 1,064 DNAs from individuals of 51 different populations, representing all continents. This panel captures most of the extant human diversity (7). To avoid ascertainment bias in this multiethnic analysis, we picked a set of SNPs within *CD28*, *CTLA4*, and *ICOS*, following several criteria: gene-centered distribution (because the aim was to analyze LD between variants at the loci rather than to define locus-wide haplotype-block structure); original identification in sizeable DNA panels (>20 individuals); and known polymorphism in at least three major geographical groups. Markers previously reported to be associated with T1D and Graves' disease (CTLA4 −318, +49, CT60) were added to this list, eventually totaling 22 SNPs

Author contributions: V.B., C.B., and D.M. designed research; V.B., M.R., and W.B. performed research; V.B., M.R., P.S., W.B., C.B., and D.M. analyzed data; and V.B., C.B., and D.M. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: LD, linkage disequilibrium; EHH, extended haplotype homozygosity; HGDP, Human Genome Diversity Panel; CEPH, Centre d'Étude du Polymorphisme Humain.

‡To whom correspondence should be addressed at: Section on Immunology and Immunogenetics, Joslin Diabetes Center, One Joslin Place, Boston, MA 02215. E-mail: cbdm@joslin.harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0610124104/DC1.

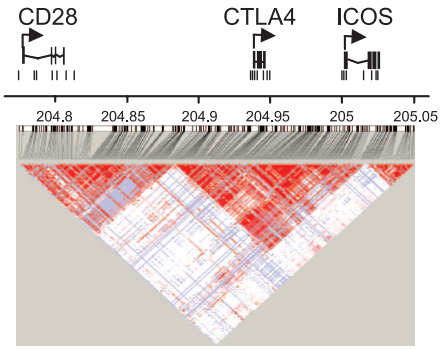

Fig. 1. The human costimulatory locus on 2q33. (*Upper*) Map of the costimulatory locus on human chromosome 2 (position in Mb, release July 2003). The position of the SNPs used in this article is indicated. (*Lower*) LD plot of the costimulatory locus in the HapMap CEU data (January 2006) by standard Haploview color code (red, $|D'| > 0.8$ and log-likelihood ratio ≥ 2; blue, $|D'| > 0.8$ and log-likelihood ratio < 2).

(Fig. 1). Genotyping was performed by fluorogenic PCR (TaqMan) [supporting information (SI) Table 1; success rate >99.6%, concordance rate 100%; complete genotype data are presented in SI Table 2]. Most SNPs were polymorphic as expected, with minor allele frequencies >5%, putting them in the range of common variants in most populations (SI Table 2). Overall, these distributions suggest that old variants were being studied, widespread among human populations despite the geographical expanse of the HGDP, thereby effectively capturing variation in very different ethnic groups.

SNP Distributions in Population Groups. To compare human populations for the representation of allelic variation in the costimulatory receptor region, we computed Wright's F_{st} statistic (8), which measures the degree of genetic differentiation between populations or population groups. The grid representation of F_{st} values computed across all SNPs (9) for each population pair (Fig. 2*A*) shows a clear differentiation between population groups. Although most F_{st} values are not significantly different from the expected genome-wide distribution (10, 11), the average F_{st} in population-to-population analysis revealed significant differences: African populations, as well as the populations of the East Asian group, clearly stand out, demonstrating higher F_{st} values when compared with all other population groups, whereas populations from the Middle East, South Asia, and Europe tend to show only weak differentiation from each other. When broken down into individual SNPs, distinctive patterns emerge (Fig. 2*B*). First, the *ICOS* coding region (*ICOSc*) shows very little differentiation between population groups, in contrast to the strong patterns at other loci. As shown below, a strong recombination hot spot is present within the first intron of *ICOS*, thus insulating the *ICOSc* markers from the remainder of the costimulatory locus. As such, *ICOSc* evolutionary trajectory would have been dissociated from *CD28*, *CTLA4*, and *ICOSp*, thus preserving it from the footprint of demographic and selective events that affected other loci. Second, although the SNPs in other blocks show strong differentiation between population groups, the exact patchwork structure varies, even for adjacent markers. For instance, *CTLA4* SNPs partition into two types (SNPs 1, 3, 5, 6, and 7 distribute very differently from SNPs 2 and 4). Finally, three SNPs stand out from others in showing a markedly higher level of differentiation. ICOSp.2 and ICOSp.4 undergo an inversion of their allele frequencies in African vs. non-African populations (SI Table 2), thus leading to elevated (>0.5) pairwise F_{st} values among populations in these respective groups. A similar phenomenon differentiates CD28.6 in East Asian and Native American populations compared with the other members of the HGDP panel. This computation of pairwise F_{st} reveals that >350 population pairs register F_{st} values >0.4 in CD28.6, ICOSp.2, and ICOSp.4. These F_{st} values appeared to be in the highest range (12), possibly suggestive of selection/adaptation events, much as extreme population differentiation at the lactase and *FY* loci denote differential selection by food or pathogen exposure (12, 13). To obtain some measure of the significance of these F_{st} distributions, we compared F_{st} profiles at these three SNPs with those of unrelated SNPs across the genome using data from 312 SNPs deposited in the HGDP–Centre d'Étude du Polymorphisme Humain (CEPH) database. The vast majority of SNPs demonstrated $F_{st} > 0.4$ in far fewer pairwise population comparisons (Fig. 2*C*). Furthermore, the magnitude of F_{st} values found in CD28.6, ICOSp.2, and ICOSp.4 also put them among the most differentiated SNPs across the HGDP (as exemplified by 98th percentile distribution) (Fig. 2*C*). Because variants typed in the HGDP panel are themselves likely to have been chosen based on strong biological/population genetic priors, CD28.6, ICOSp.2, and ICOSp.4 clearly stand out from genome-wide distribution by showing unusual variation between population groups.

To confirm the peculiar pattern of population differentiation

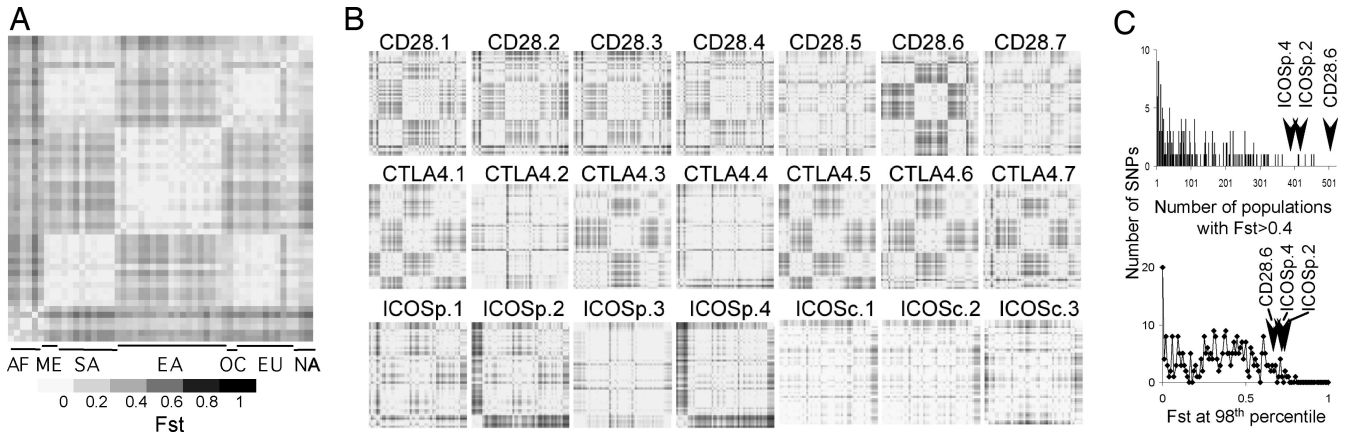

Fig. 2. F_{st} values in the costimulatory locus. (*A*) Overall pairwise F_{st} values between populations computed across all SNPs. Population groups: Africa (AF), Middle East (ME), South Asia (SA), East Asia (EA), Oceania (OC), Europe (EU), and Native American (NA). The representation is symmetric. (*B*) F_{st} plots for individual SNPs in the locus. (*C Upper*) Pairwise F_{st} distribution in 312 SNPs deposited in the HGDP database. The graph represents the number of SNPs (*y* axis) with a given number of pairwise F_{st} values >0.4. (*C Lower*) Distribution of 98th percentile values of the SNPs represented in *Upper*.

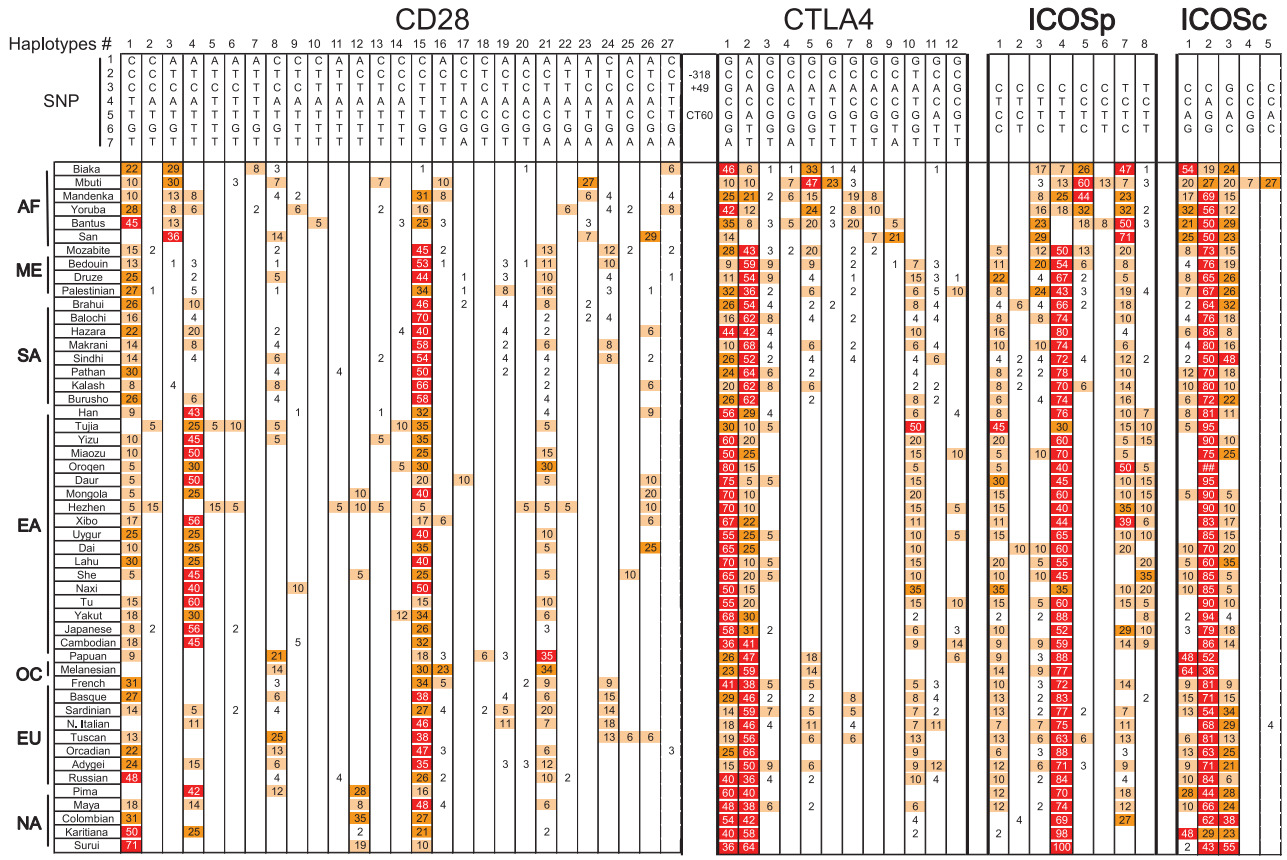

Fig. 3. Shared haplotypes at each locus. (*A*) Haplotypes were reconstructed computationally in each population for SNPs in *CD28*, *CTLA4*, *ICOSp*, and *ICOSc*. Shared haplotypes present in at least two populations and at 5% or more in one population are represented. Population groups are as in Fig. 2. Cells are color-coded based on the frequency of that haplotype in a given population (ochre, 5–20%; orange, 20–35%; red, >35%).

at these three SNPs, we applied the analysis of molecular variance method, which rests on the analysis of the components of covariance of allele frequencies within populations, or within population groups, or worldwide. Published observations show that most of the variation lies within individual populations (80–90%), with only 5–10% being contributed by diversity between regions (14). Here, analysis of molecular variance revealed stronger interregional differences in CD28.6, ICOSp.2, and ICOSp.4, with ≈20% or more of the allele frequency variation being at the interregional level (SI Table 3). These data strongly suggest that major demographic/selective events affected the costimulatory locus in human populations after the out-of-Africa migrations.

Haplotypes at Each Locus. To determine how these single variants grouped into haplotypes and how these haplotypes are distributed worldwide, we reconstructed haplotypes at *CD28*, *CTLA4*, *ICOSp*, and *ICOSc* for each of the HGDP populations. A number of recurrent haplotypes were identified (>5% frequency in at least two populations) (Fig. 3). These recurrent haplotypes together account for a major proportion of the gene pool in all populations (96.9% or more on average). A recent study systematically interrogated a number of SNPs in the *CTLA4* region in the same CEPH-HGD panel (15), leading to the description of 10 haplotypes, which agree well with the ones we identified.

In almost all populations, the distribution of haplotypes includes one dominant representative that accounts for approximately half of the chromosomes, and one or a few minor ones that complete the pool, as commonly observed in other genes and expected from theoretical considerations (16). Differential SNP distributions translate here as a distribution of haplotypes that is markedly different between the various population groups, with delineations that correspond precisely to geographical and historical boundaries. For example, East Asian populations show a dominance of CD28.h4, which is rare or absent in other regions, where CD28.h15 accounts for 46% (26–70%) of chromosomes, except in Africa, where CD28.h1 and h3 tend to predominate. This contrasting frequency distribution of major haplotypes mirrors their distance at the sequence level: haplotypes CD28.h15 and CD28.h4 are very distant from each other, differing at five of seven positions. *CTLA4* haplotypes reveal a similar picture, with two major haplotypes (CTLA4.h1 and CTLA4.h2) of diametrically opposite distribution and sequence composition.

For the *ICOSp* block the situation is different. The same haplotype (ICOSp.h4) dominates, except in Africa, where ICOSp.h5 and ICOSp.h7 predominate; interestingly, these two haplotypes contribute alleles at the ICOSp.2 and ICOSp.4 SNPs that showed such strong differentiation. Thus, the haplotypic composition at *CD28* and *CTLA4* brings out a strongly differentiated East Asian pattern, whereas the *ICOSp* block seems to isolate a specifically African profile.

Gene–Gene LD Patterns and Haplotypes. We then analyzed patterns of LD across the region. *CD28*, *CTLA4*, and *ICOS* have opposite effects in controlling the expansion and differentiation of T lymphocytes during the immune response, and pressure to maintain complementary expression and function of these molecules might leave a signature in the long-range LD in the region. From another angle, shared transmission of variants, potentially

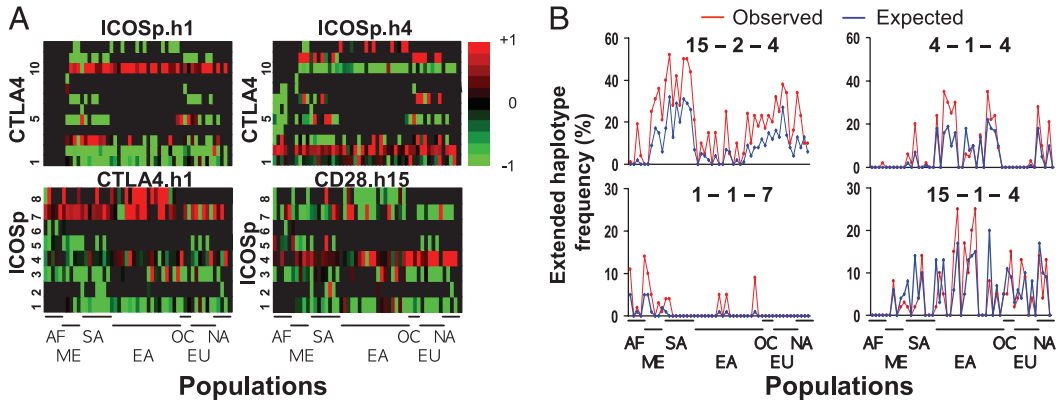

Fig. 4. Extended haplotypes account for a large fraction of the variation in the costimulatory locus. (A) Examples of pairwise association between haplotypes at *CD28* or *CTLA4* and *ICOSp* across the HGDP. Each matrix depicts the strength of the LD among individual haplotypes at *ICOSp* and *CTLA4* (*y* axis, *Upper*) or *CD28*/*CTLA4* and *ICOSp* (*y* axis, *Lower*) in each population (*x* axis). Signed D′ is used as a metric for LD, where overrepresented associations are shown in shades of red, underrepresented associations are shown in shades of green, and neutral/absent combinations are shown in black. Haplotypes at each locus are numbered as defined in Fig. 3. (*B*) Frequency (%) of whole region extended haplotypes at *CD28*, *CTLA4*, and *ICOSp* as reconstructed with Phase. The predicted frequency is the product of individual haplotype frequencies in each population (assuming no LD among haplotypes at *CD28*, *CTLA4*, and *ICOSp*).

differing between populations, would need to be accounted for in association studies.

Previous studies in Caucasians revealed a pattern characterized by strong LD around *CD28*, a second block encompassing *CTLA4*, and *ICOSp* itself separated from a third block by an intense recombination hot spot located in the first intron of *ICOS*, as exemplified by data from the CEU (CEPH of European descent) population in the HapMap (Fig. 1) (17, 18).

To avoid artifactual admixture effects, we studied the patterns of LD in each population group using two different metrics, |D′| and log-likelihood ratio. Complex and somewhat diverse LD patterns could be observed in different population groups (SI Fig. 6). The overall LD pattern previously described in Caucasians was recapitulated, in particular the recombination hot spot in the first intron of *ICOS*. Very significant LD was observed between markers at the three genes, but with significant differences between geographical groups. As expected, LD was less marked in the African group (19), and a west–east gradient could be observed in the span and strength of LD across the region, with Far East Asian populations showing the highest amount of LD. In all cases, significant LD was present between the genes, extending through and beyond regions where local LD had decayed (for instance, LD from proximal *CD28* markers can be found to reach to *ICOSp* markers, beyond several *CD28* and *CTLA4* SNPs with which they share little linkage).

Given this LD between costimulatory receptor family members, we then asked how haplotypes defined at each locus are arranged together and whether there are common extended haplotypes that associate haplotypes at each of the genes. To search for such chromosomal combinations, genotypes for all *CD28*, *CTLA4*, and *ICOSp* SNPs were phased within each population with Phase2.1 (20–22). The data were analyzed for the presence of overrepresented or underrepresented haplotype pairs relative to random association of their components, as ascertained by computing a multiallelic D′. Fig. 4*A* illustrates the results for different ICOSp haplotypes and their association with *CTLA4* or *CD28* haplotypes. Many *CTLA4* or *CD28* haplotypes showed no preferential linkage to *ICOSp* haplotypes, or only scattered signals that are likely to represent fluctuations specific to single populations. For several other combinations, however, the departure from expected frequencies was seen across a broad swath of populations. Strikingly, ICOSp.h4 showed recurrent associations with a number of *CTLA4* and *CD28* haplotypes, in particular CTLA4.h2 throughout, CTLA4.h11 in "West of the Himalayas" populations, and CTLA4.h12 in East Asians. Similarly, other positive association are observed throughout (e.g., CTLA4.h1 and ICOSp.h7, or ICOSp.h1 and CTLA4.h10), whereas other combinations appear to be underrepresented. These nonrandom associations between haplotypes extended to three-gene combinations (Fig. 4*B* and SI Table 4). Some of these extended haplotypes proved quite frequent, such as the combination of CD28.h15, CTLA4.h2, and ICOSp.h4 (abbreviated as 15-2-4), which accounts for 20–60% of the chromosomes in the West of the Himalayas populations, a frequency substantially higher than expected in the absence of LD. Similar patterns are observed for the 4-1-4 combination in East Asian populations and the 3-1-7 and 1-1-7 combinations in African populations (SI Table 4 and data not shown). Counts of homozygous and heterozygous individuals for the major extended haplotypes (15-2-4 and 4-1-4) were in Hardy–Weinberg equilibrium, thus confirming their reality beyond statistical reconstruction.

Extended Haplotype Analysis Across the Locus. These extended haplotypes involving haplotypes at all three loci dominate European and Asian populations while being rare in African groups, raising the issue of their origin. Bottlenecks, population structure, and selection can each leave a distinctive signature on the haplotype structure of a region. In the case of a bottlenecked population, the overall diversity is expected to be reduced, with a few haplotypes accounting for the bulk of the chromosomes. Recent positive selection (20,000 years or less), differentially affecting population groups, is characterized by a pattern of high-frequency haplotypes maintaining long neighboring segments in strong LD (23). To explore the costimulatory locus for such signatures, we analyzed the LD patterns using extended haplotype homozygosity (EHH) statistics, which focus on the relationship between the frequency of a given core haplotype and the span of SNP homozygosity at increasing distances from that core (24). More frequent core haplotypes, such as CTLA4.h2 in West of the Himalaya populations, demonstrated longer spans of LD than other core haplotypes, which showed a much more branched structure (e.g., CTLA4.h1 in Caucasians) (Fig. 5*A*). The presence of decaying haplotypes around CTLA4.h1 in West of the Himalayas populations also strongly argues against a significant bottleneck, which would have been expected to erase most of the preexisting variation and affect similarly the LD around both haplotypes. The converse was true in East Asian populations, where more persistent LD was associated with the dominant CTLA4.h1 than around CTLA4.h2. Interestingly, common extended haplotypes repre-

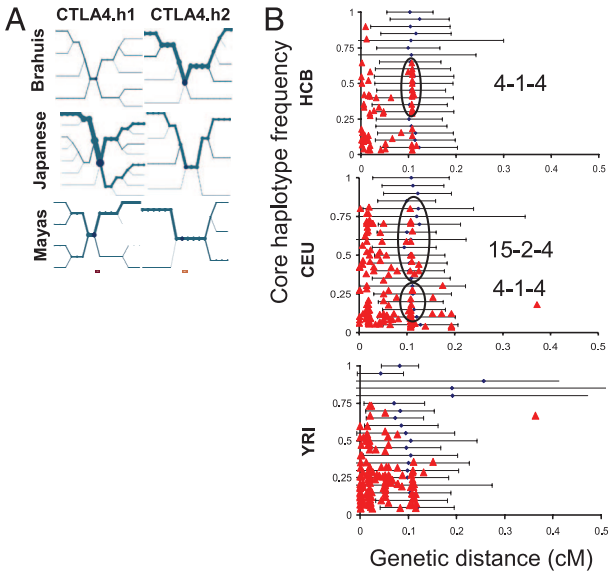

Fig. 5. EHH across the costimulatory locus. (*A*) Extended haplotypes bearing core CTLA4.h1 or CTLA4.h2 were reconstructed in West of the Himalayas and East Asian populations. Each drop in haplotype homozygosity is depicted as a bifurcation. Branches' width is proportional to the frequency of haplotypes still homozygous at a given position. (*B*) Span of core haplotypes with EHH > 0.8 found in the HGDP panel relative to their frequency. Core haplotypes embedded in major extended haplotypes are highlighted. For comparison, matched HapMap chromosome 2 data are displayed as mean ± standard deviation [CEU, CEPH (Utah residents with ancestry from northern and western Europe); YRI, Yoruba in Ibadan, Nigeria; HCB, Han Chinese in Beijing].

sent much of the variation observed in Native American populations, thus serving as an internal control for the test's sensitivity to bottlenecks.

These data suggest that widespread bottleneck effects are unlikely to entirely account for the presence and frequency of the 15-2-4 and 4-1-4 extended haplotypes. To assess the significance of these extended haplotypes in a genome-wide context, we compared the core haplotypes found in the HGDP populations with reference data from the HapMap project (11), plotting the LD span around core haplotypes against their frequency (Fig. 5*B*). The extended haplotypes of the costimulatory region partitioned in two distinct groups: some showed rapid decay, with short lengths (0.05 cM or less). On the other hand, core haplotypes embedded in the major extended haplotypes (15-2-4 and 4-1-4) stood out, spanning much of the overall length of the costimulatory locus. To obtain a better sense of the chromosome-wide distribution of core haplotype length and frequency, 57,500 SNPs matching our initial selection criteria were picked across the HapMap data for *Chr2*. After matching for genetic distance, these SNPs resulted in 1,567 regions, for which haplotypes were reconstructed and processed through Sweep. The inverse relationship between LD span and frequency depicted in Fig. 5*B* follows the expected pattern under neutral evolution (more clearly observed in the data from African populations, where the overall LD is less). The monotonic haplotype length (≈0.11 cM) across the frequency spectrum found in non-African populations most probably results from the ascertainment scheme of SNPs, conditioned to match the original genetic distances across the costimulatory locus. Over this chromosome-wide backdrop, 15-2-4-related or 4-1-4-related cores fall on the mean of the distribution, whereas other cores tend to show shorter than average LD spans.

These data suggest that the costimulatory region does reveal evidence of stronger long-range LD in some of its variants, which nevertheless do not stand out in a *Chr2*-wide fashion.

Discussion

This worldwide analysis of the distribution of polymorphisms at the three linked costimulatory loci resulted in two important conclusions. First, there is significant differentiation between populations in the representation of alleles or haplotypes at the *ICOS* promoter, with the emergence of frequent variants in postmigration populations that are rare in African populations. Second, extended haplotypes encompassing variants at *CD28*, *CTLA4*, and *ICOSp* represent an important fraction of chromosomes. These two conclusions are connected, in that the extended haplotypes that dominate in postmigration populations are those that carry the expanded variants and have significant implications for genes' coordinated function in regulating immune and autoimmune responses.

Extended Haplotypes and ICOSp Variants. The original demonstrations of long-range LD in the MHC were particular in showing patterns of LD resurgence, linking loci over long genomic distances in which LD decayed (25). The patterns were interpreted to reflect preferential functional associations between allelic variants from ancestral combinations generated by recombination and then fixed and maintained by selective pressure (25, 26). Here the extended haplotypes that span the costimulatory region occur on a smaller genomic scale, such that the LD observed, even though it does show patterns of resurgence after LD decay, would also be compatible with simpler origins such as limited recombination across extended blocs.

Where do costimulatory extended haplotypes originate? Various mechanisms could result in such extended haplotype formation and stabilization, including selection of optimal allele combinations (as exemplified by the MHC locus) and ancestral bottleneck events. Neutral combinations of alleles arising in the wave front of human range expansions could have drifted to their current frequency (27). Despite apparent breakdowns in LD leading to haplotype blocks, extended haplotypes spanning more than one LD block are deemed to be common in the human genome because of the low recombination frequency of even the hottest hot spots (28). Importantly, EHH analyses show these major extended haplotypes to be relatively longer than for other haplotype cores, suggesting possible selective events. It is probably not a coincidence that the extended haplotypes that dominate postmigration populations (15-2-4 West of Himalayas and 4-1-4 in East Asia) share the ICOSp.h4 haplotype. This high frequency of ICOSp.h4 is largely responsible for the allele frequency reversal of ICOSp.2 and ICOSp.4 between African and non-African populations. Comparison to other data from the HGDP places the degree of population differentiation at these SNPs at the extreme of the F_{st} distribution. One can envision that new selective influences, appearing after the migration out of Africa, would operate on the extended haplotypes that carried ICOSp.h4 in a subdivided ancestral population (29). Human populations migrating from the predominantly tropical environment of Africa to quite different climates would have faced a very different range of environmental challenges, with a reduced parasitic load at increasing latitudes (30). One may speculate that signatures of adaptation to these varying challenges would be reflected in the pattern of variability at the costimulatory receptor region, because these genes control the intensity and "flavor" of immune responses. Recent functional studies incriminated SNPs in strong LD ($r^2 \approx 0.8$) with ICOSp.2 and ICOSp.4 in the susceptibility to asthma and the production of Th2 cytokines (31). Alleles corresponding to the common African variants were associated with higher Th2 cytokines titers, i.e., those particularly associated with response to parasites. One could envision continuous selective pressure in Afri-

can populations leading to the predominance of haplotypes ICOSp.h5, ICOSp.h6, and ICOSp.h7 (>60%), all bearing ICOSp.2C, whereas the migration out of Africa favored a gradual rise in frequency of ICOSp.h4 and the derived haplotypes ICOSp.h1 and ICOSp.h8. Because these events are expected to have happened over the last 60,000–80,000 years, selective signatures such as EHH over long stretches would be more modest than the selective signature of malaria or cattle domestication, both of which are more recent (12, 23, 24, 32).

Implications for Immunity and Autoimmunity. In West of the Himalayas populations the high rise in frequency of the 15-2-4 extended haplotypes leads to an inversion of CT60 (CTLA4.5) allele frequencies. The CT60 polymorphism, located in the 3′ UTR of *CTLA4*, was the SNP most tightly associated with Graves' thyroiditis (and type 1 diabetes to a lesser extent) (3). In Caucasian populations 15-2-4 is the major haplotype tagged by the protective CT60 A allele (as well as other high-ranking SNPs in ref. 17), and one may question whether a global haplotype protective effect is conferred by 15-2-4 and related extended haplotypes, rather than by single polymorphisms such as CT60. For instance, a recent investigation of susceptibility to celiac disease concluded that an extended haplotype stretching over *CD28*, *CTLA4*, and *ICOSp* was significantly overrepresented in controls and presumably more protective than any one SNP (33). Interestingly, this haplotype partially overlaps with the 15-2-4 haplotype defined here. The same haplotype imputed in Graves' data from Ueda *et al.* (17) also reveals a higher protection than that conferred by a CT60 A genotype alone (OR = 0.64 for CT60A, $P = 1.6 \times 10^{-6}$; OR = 0.58 for Brophy's haplotype 8, $P = 1.9 \times 10^{-7}$).

Based on the data presented here, functional analyses taking into account extended haplotypes across the costimulatory locus are needed to further dissect the relative role of individual variants or their combinations.

Materials and Methods

DNA Samples. DNA samples were obtained from a worldwide DNA panel from the CEPH (Paris, France; HGDP1-2-1) (6).

SNP Genotyping. Most genotypes were determined by allele-specific fluorogenic PCR (detailed in SI Table 1).

Bioinformatics Analysis. Computational techniques are described in *SI Materials and Methods*. Most classical population genetics tests were run by using Arlequin2.1. Populations were grouped based on their geographical clustering, without prior inference about their genetic similarities. F_{st} results were further processed in S+. Phase2.1 was used to reconstruct haplotypes in individual populations (20, 22) and run three times with different random seeds, and the results were checked for concordance. Patterns of LD were calculated and displayed by using Haploblockfinder (34). Haplotypes were reconstructed across the whole costimulatory locus, one population at a time, using Phase2.1, and the major haplotypes at *CD28*, *CTLA4*, and *ICOS* were extracted from the computed extended haplotypes. For dual-locus plots, a signed D′ metric was applied on each haplotype combination observed in the same data set, and log-likelihood ratios for significance of linkage were computed for all haplotype pairs across all populations. EHH tests were carried out with the Sweep package (24). For analysis across chromosome 2, phased HapMap data were used; SNPs matching this study's ascertainment criteria were selected, and haplotypes were reconstructed and processed with the EHHCorrelations batch function (maximum number of SNPs in core = 10, EHH to match 0.8). For F_{st} analysis in the CEPH-HGD panel, SNP data from the CEPH web site (www.cephb.fr/hgdp-cephdb) were analyzed by using an S+ implementation of Arlequin's algorithm.

We thank Drs. A. Sharpe, J. Wakeley, D. Altshuler, A. Sanchez-Mazas, and L. Excoffier for inspiring discussion and H. Ranu for help with the robotics. This work was funded by National Institutes of Health Grant P01-AI056299, the William T. Young Chairs in Diabetes Research, and the Joslin Diabetes and Endocrinology Research Center funded cores.

1. Greenwald RJ, Freeman GJ, Sharpe AH (2005) *Annu Rev Immunol* 23:515–548.
2. Teft WA, Kirchhof MG, Madrenas J (2006) *Annu Rev Immunol* 24:65–97.
3. Gough SC, Walker LS, Sansom DM (2005) *Immunol Rev* 204:102–115.
4. Greve B, Vijayakrishnan L, Kubal A, Sobel RA, Peterson LB, Wicker LS, Kuchroo VK (2004) *J Immunol* 173:157–163.
5. Abecasis GR, Ghosh D, Nichols TE (2005) *Hum Hered* 59:118–124.
6. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, *et al.* (2002) *Science* 296:261–262.
7. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) *Science* 298:2381–2385.
8. Wright S (1965) *Evolution (Lawrence, Kans.)* 19:395–420.
9. Schneider S, Roessli D, Excoffier L (2000) *Arlequin: A Software for Population Genetics Data Analysis* (Genetics and Biometry Lab, Dept of Anthropol, Univ of Geneva, Geneva).
10. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) *Genome Res* 12:1805–1814.
11. International HapMap Consortium (2005) *Nature* 437:1299–1320.
12. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) *Am J Hum Genet* 74:1111–1120.
13. Hamblin MT, Di Rienzo A (2000) *Am J Hum Genet* 66:1669–1679.
14. Jorde LB, Wooding SP (2004) *Nat Genet* 36:S28–S33.
15. Ramirez-Soriano A, Lao O, Soldevila M, Calafell F, Bertranpetit J, Comas D (2005) *Genes Immun* 6:646–657.
16. Innan H, Zhang K, Marjoram P, Tavare S, Rosenberg NA (2005) *Genetics* 169:1763–1777.
17. Ueda H, Howson JMM, Esposito L, Heward J, Snook H, Chamberlain G, Rainbow DB, Hunter KMD, Smith AN, Di Genova G, *et al.* (2003) *Nature* 423:506–511.
18. Haimila K, Smedberg T, Mustalahti K, Maki M, Partanen J, Holopainen P (2004) *Genes Immun* 5:85–92.
19. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, *et al.* (2002) *Science* 296:2225–2229.
20. Stephens M, Smith NJ, Donnelly P (2001) *Am J Hum Genet* 68:978–989.
21. Stephens M, Donnelly P (2003) *Am J Hum Genet* 73:1162–1169.
22. Stephens M, Scheet P (2005) *Am J Hum Genet* 76:449–462.
23. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES (2006) *Science* 312:1614–1620.
24. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, *et al.* (2002) *Nature* 419:832–837.
25. Yunis EJ, Larsen CE, Fernandez-Vina M, Awdeh ZL, Romero T, Hansen JA, Alper CA (2003) *Tissue Antigens* 62:1–20.
26. Traherne JA, Horton R, Roberts AN, Miretti MM, Hurles ME, Stewart CA, Ashurst JL, Atrazhev AM, Coggill P, Palmer S, *et al.* (2006) *PLoS Genet* 2:e9.
27. Edmonds CA, Lillie AS, Cavalli-Sforza LL (2004) *Proc Natl Acad Sci USA* 101:975–979.
28. McVean G, Spencer CC, Chaix R (2005) *PLoS Genet* 1:e54.
29. Di Rienzo A, Hudson RR (2005) *Trends Genet* 21:596–601.
30. Guernier V, Hochberg ME, Guegan JF (2004) *PLoS Biol* 2:e141.
31. Shilling RA, Pinto JM, Decker DC, Schneider DH, Bandukwala HS, Schneider JR, Camoretti-Mercado B, Ober C, Sperling AI (2005) *J Immunol* 175:2061–2065.
32. Beja-Pereira A, Luikart G, England PR, Bradley DG, Jann OC, Bertorelle G, Chamberlain AT, Nunes TP, Metodiev S, Ferrand N, *et al.* (2003) *Nat Genet* 35:311–313.
33. Brophy K, Ryan AW, Thornton JM, Abuzakouk M, Fitzgerald AP, McLoughlin RM, O'Morain C, Kennedy NP, Stevens FM, Feighery C, *et al.* (2006) *Genes Immun* 7:19–26.
34. Zhang K, Jin L (2003) *Bioinformatics* 19:1300–1301.