A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection


1Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. 2Broad Institute of MIT and Harvard, Cambridge, MA, USA. 3Mount Sinai School of Medicine, New York, NY, USA. 4Department of Systems Biology, Harvard Medical School, Boston, MA, USA. 5Department of Biology, MIT, Cambridge, MA, USA.

*These authors contributed equally to this work. †To whom correspondence should be addressed. E-mail: psabeti@oeb.harvard.edu (P.C.S.), shari.grossman@post.harvard.edu (S.R.G), ilya_shl@alum.mit.edu (I.S.)

The human genome contains hundreds of regions whose patterns of genetic variation indicate recent positive natural selection, yet for most the underlying gene and the advantageous mutation remain unknown. We developed a method, Composite of Multiple Signals (CMS), that combines tests for multiple signals of selection and increases resolution by up to 100-fold. Applying CMS to candidate regions from the International Haplotype Map, we localized population-specific selective signals to 55 kb (median), identifying known and novel causal variants. CMS can identify not just individual loci but implicates precise variants selected by evolution.

Numerous methods have been developed to exploit signatures left by positive natural selection to identify genomic regions in the human genome harboring recent local adaptations, presumably to such pressures as infectious disease, changes in diet, and new environments (1, 2). Hundreds of such regions have been identified, but they are typically large, 100s of kilobases to megabases, and contain many genes and thousands of polymorphisms. In only a handful has there been much progress in identifying the causal mutations and extracting these biological insights about their function. More powerful methods are needed to pinpoint the exact mutations driving evolution, especially as increasingly powerful sequencing technologies make it possible to sequence the genomes of humans and many other species.

Initial surveys of selective events have relied on three patterns of variation caused by a new beneficial mutation rising quickly in prevalence in a population: (1) Long Haplotypes: An allele under positive selection increases in frequency so rapidly that long-range associations with neighboring polymorphisms – the “long-range haplotype” – are not disrupted by recombination. (2) High frequency derived alleles: A new (non-ancestral, or derived) allele rises to a frequency higher than expected under genetic drift, carrying neighboring derived alleles with it. (3) Highly differentiated alleles: Positive selection in one geographic region causes larger frequency differences between populations than for neutrally evolving alleles. In humans, these three signals are detectable back to between 30 to 80 thousand years ago (2).

If each signature provides distinct information about selective sweeps, combining the signals should have greater power for localizing the source of selection than any single test. As inputs to a composite statistic we chose two established metrics for haplotype length (iHS and XP-EHH) and one for population differentiation (FST). We also developed and incorporated two additional tests. ΔDAF tests for derived alleles that are at high frequency relative to other populations; it is more sensitive for distinguishing selected alleles than the simple derived allele frequency (DAF, fig. S1). ΔiHH measures the absolute rather than the relative length of haplotypes, and is particularly sensitive for identifying lower frequency selected alleles.

To characterize each test’s ability to localize signals of recent local adaptation spatially and to distinguish causal variants from nearby neutral markers, we simulated neutrally-evolving regions and regions containing a positively selected allele by standard coalescent approaches (6). We tested a range of demographic models, including a standard neutral model, a calibrated model of European, East Asian, and West African populations and several more extreme models. Regions under selection were modeled as containing a single, centrally located selected variant that appeared within the last 5000 – 30,000 thousand years, was subject to a specified
intensity of selection, and rose to present-day frequencies ranging from 20% to 100% (table S1).

For each model set we generated 1500 replicates, each consisting of 1Mb of simulated sequence data (~10,000 polymorphisms) for 120 chromosomes from each populations. In addition, we generated a dataset that matched the frequency distribution and density of Phase II of the International Haplotype Map Project (HapMapII). (7)

Under all scenarios, each of the five statistics had distinguishable distributions for causal and for neutral variants (including neutral variants in selected regions). The FST and XP-EHH signals peaked more narrowly around the causal variant, making them useful for spatial localization, but poorly distinguished the precise causal variant (Fig. 1, fig. S2). In contrast, iHS, ΔHiHH and ΔDAF contributed little to spatial resolution, but better distinguished causal variants. The five tests were nearly uncorrelated in neutral regions, and only weakly correlated for neutral variants within selected regions (fig. S3). In the latter case, correlation was appreciable only immediately around the causal variant.

As each of the five tests had power to distinguish selected from non-selected variants, and were only weakly correlated for neutral variants, we combined them in a composite likelihood statistic, termed the Composite of Multiple Signals (CMS). For each test i, we estimated from simulation the probability P of a score $s_i$ if selected and if unselected. Assuming a uniform prior probability of selection $\pi$, the CMS score is the approximate posterior probability that the variant is selected:

$$CMS = \prod_{i=1}^{n} \left( \frac{P(s_i | selected) \times \pi}{P(s_i | selected) \times \pi + P(s_i | unselected) \times (1-\pi)} \right)$$

We calculate the CMS score and significance (on the basis of the genome-wide distribution of scores) for every variant. To localize a signal, the distribution of CMS scores across the entire region is used to estimate a posterior probability curve for the position of the causal variant and determine 90% credible intervals (SOM).

In simulations, CMS showed power both to localize the selection signal spatially and distinguish the causal variant (Fig. 1, K and L). While single tests provided weak localization (~1 Mb), CMS localized the signal to average 89kb (for full sequence data) and contained the causal variant in 90% of cases (Fig. 3, A and B). With sparser genotype data (corresponding to HapMapII), CMS localized to 104 kb, even when the causal variant was absent from the dataset. CMS also showed greater specificity for the causal variant. At score thresholds giving 90% power to detect the true causal variant, the individual tests identified ~500-1500 candidate causal variants per region, while CMS narrowed the signal to ~100 (table S2). The causal variant was among the top twenty variants in half of cases and was the highest scoring variant in a quarter of cases, with high power given that we included

The CMS results were robust under all demographic scenarios tested (constant population size and bottlenecks of varying strengths), even though the test was optimized for a single model (6) (fig. S4). The most extreme bottleneck scenarios did increase the number of high scoring variants in neutral regions, but the false positive rate remained below 0.01% in all cases (SOM, (8)). These false positives occurred as isolated points, easily distinguishable from the clearly-defined peaks found in selected regions (table S4).

We then applied CMS to empirical human data for 185 candidate regions identified as under recent positive selection in HapMapII data. The dataset includes 3.1 million variants genotyped in three populations: Northern Europeans, West Africans (Yoruba from Nigeria), and East Asians (Chinese and Japanese) (7).

As positive controls, we examined several well-characterized regions under positive selection (Figs. 2 and 3). In three regions (containing, respectively, SLC24A4, LCT, and EDAR), a putative causative variant has been previously identified and genotyped in HapMapII (2, 3). In each region, the variant was within the top ten CMS scores, out of 1000 - 1500 variants in the region. We also examined four regions (350kb-1MB) containing pigmentation-related genes (MATP, TYRP1, OCA2 and HERC2, and KITLG) that are suggested targets of recent selection, but where no candidate variant has been proposed (1, 9, 10). CMS improved the spatial resolution by 3-80 fold, and, in each case, the narrowed region contains a single pigmentation-related gene. In each case, a strong CMS signal is found at a variant known to be associated in the human population with eye color or skin pigmentation (9).

We then examined the remaining 178 candidate HapMapII regions, containing ~1500 genes, for which the selected locus and variant are unknown. After application of CMS, 64 regions contained a single gene, 35 contained multiple genes, and 79 contained no genes at all. CMS suggested numerous intriguing coding and regulatory functional candidates (figs. S5 and S6; table S5).

Many regions include striking amino acid changes (table S6). For example, CMS localized a region on chromosome 10 with evidence for selection in East Asians, to the protocadherin gene PCDH15. The third highest ranking variant is an acidic-to-nonpolar (Asp-435-Ala) mutation altering a highly conserved residue predicted to lie in the Ca" binding site at the interface of cadherin repeats in the protein’s extracellular domain (SOM, Fig. 4A, figs. S7 and S8, (11)). PCDH15 plays a role in development of inner ear hair cells and maintaining retinal photoreceptors (12, 13).
Another signal in East Asians localized to the leptin receptor, LEPR. The highest-scoring variant is a Lys-109-Arg change in LEPR associated with blood pressure, glucose response, and body mass index (14).

Many signals, however, are localized to intergenic regions or regulatory changes in gene regions, suggesting that selected variants may lie in regulatory elements (which also harbor many variants affected in complex diseases). For example, a signal of selection in West Africans localized to a single gene, PAWR. Several high-scoring variants show strong association with PAWR expression uniquely in West Africans, and with no other genes in the region (fig. S9).

Another signal in West Africans localized to a 22 kb region containing two genes, USF1 and ARHGAP30. Several high-scoring SNPs in USF1 show strong association with USF1 expression uniquely in West Africans. One variant lies within an experimentally determined transcription factor binding site (15).

Beyond identifying individual gene and polymorphism targets, by reducing the number of genes within each region from ~8 to ~1 the method reveals instances of multiple genes or regulatory features of large sets of genes (16). With such data emerging for humans and other species, it will be possible to examine nearly every variant in the genome in many individuals and populations. With such data emerging for humans and other species, it may be possible to observe much of evolution’s most recent handiwork and identify many of the functional adaptations that work to shape species.

References and Notes
16. J. Reiners et al., Human molecular genetics 14, 3933 (Dec 15, 2005).
17. P. D. Thomas et al., Nucleic acids research 31, 334 (Jan 1, 2003).
19. We thank K. Andersen, J. Lohmueller, B. Stranger, S. McCarroll, K. Lohmueller, K. Lindblad-Toh, M. Guttmann, and J. Rinn for functional guidance, and E. Phelan, D. Altschuler, and the Sabeti Lab for helpful discussions throughout. PCS is supported by the Burroughs Wellcome and Packard Foundations. EKK is supported by the American Cancer Society.

Supporting Online Material
www.sciencemag.org/cgi/content/full/science.1183863/DC1
Methods
Figs. S1 to S10
Tables S1 to S7
References
Fig. 1. CMS localizes selection and identifies causal variants better than single tests. Left: top (red) and bottom (black) 5% of scores and mean score (black, dashed) in 1MB surrounding causal mutation (located at red dashed line). Right: distribution of scores for the causal variant (red), nearby unselected variants (blue) and variants in regions without selection (grey, below axis). The composite test (CMS), outperforms individual tests for (K) localizing the selective signal and (L) distinguishing the causal variant.

Fig. 2. Localizing selection at MATP. Scores of six individual tests (A to F) and CMS (G) for a region containing MATP. A non-synonymous SNP (rs16891982, F374L, red dotted line) associated with pigmentation is believed to be the mutation under selection.

Fig. 3. CMS localizes selection and identifies causal variants in simulated and empirical data. CMS analysis of: (A) simulated full sequence and HapMapII-density genotype datasets; and HapMapII selective sweeps at the genes (B) EDAR, (C) LCT, (D) SLC24A5, (E) OCA2/HERC2, (F) TYRP1, (G) KITLG. Bars on x-axis indicate genes, red bar indicates putative selected gene, blue dots show CMS values, red stars indicate putative causal alleles, red circle on LCT is a variant in a conserved transcription factor-binding motif.

Fig. 4. Coding and regulatory mutations identified by CMS. (A) CMS scores around PCDH15 (HapMapII data). (Red circle: non-synonymous mutation (D435A).) (B) Homology modeling of the PCDH15 cadherin-4 domain (red) predicts that D435A (red rods) is among the residues (blue) coordinating calcium ions (green) essential to cell-cell adhesion. (C and D) Variants identified by CMS involved in gene regulation. Upper: CMS scores for each HapMapII SNP within region originally identified as under selection. Lower: strength of association in West African samples between genotype and gene expression level for PAWR (C) and USF1 (D).