References and Notes

1. C. H. Pui, W. E. Evans, *N. Engl. J. Med.* **354**, 166 (2006).
2. I. Aifantis, E. Raetz, S. Buonamici, *Nat. Rev. Immunol.* **8**, 380 (2008).
3. P. Van Vlierberghe *et al.*, *Blood* **108**, 3520 (2006).
4. T. H. Rabbitts, *Genes Dev.* **12**, 2651 (1998).
5. C. H. Nam, T. H. Rabbitts, *Mol. Ther.* **13**, 15 (2006).
6. S. Hacein-Bey-Abina *et al.*, *Science* **302**, 415 (2003).
7. M. P. McCormack, T. H. Rabbitts, *N. Engl. J. Med.* **350**, 913 (2004).
8. S. J. Howe *et al.*, *J. Clin. Invest.* **118**, 3143 (2008).
9. S. Hacein-Bey-Abina *et al.*, *J. Clin. Invest.* **118**, 3132 (2008).
10. R. C. Larson *et al.*, *Oncogene* **9**, 3675 (1994).
11. R. C. Larson, H. Osada, T. A. Larson, I. Lavenir, T. H. Rabbitts, *Oncogene* **11**, 853 (1995).
12. Materials and methods are available as supporting material on *Science* Online.
13. M. A. Hall *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 992 (2003).
14. B. A. Schwarz, A. Bhandoola, *Immunol. Rev.* **209**, 47 (2006).
15. A. J. Warren *et al.*, *Cell* **78**, 45 (1994).
16. Y. Yamada *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 3890 (1998).
17. L. J. Patterson *et al.*, *Blood* **109**, 2389 (2007).
18. M. P. McCormack, A. Forster, L. Drynan, R. Pannell, T. H. Rabbitts, *Mol. Cell. Biol.* **23**, 9003 (2003).
19. A. A. Ferrando *et al.*, *Cancer Cell* **1**, 75 (2002).
20. We thank C. Talora and I. Screpanti for *Lck-Notch3-IC* transgenic mice; D. Izon for the MIG-Hhex retroviral vector; A. Strasser for comments on the manuscript; G. Smyth for advice on bioinformatics; and L. Ta, R. Bowyer, L. Mizhiritzky, and J. Davis for animal husbandry. This work was supported by grants (to M.P.M., S.M.J., and D.J.C.) from the Cancer Council of Victoria and National Health and Medical Research Council of Australia (NHMRC) (project grant 382901); a grant-in-aid (to M.P.M. and D.J.C.) from the Leukaemia Foundation; and grants (to T.H.R.) from the Medical Research Council UK (programme grant G0600914) and Leukaemia Research UK (programme grant 07036). D.J.C. is an R.D. Wright Biomedical Research Fellow, and S.M.J. is a Principal Research Fellow of the NHMRC.

Supporting Online Material

www.sciencemag.org/cgi/content/full/science.1182378/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 and S2
References

23 September 2009; accepted 23 December 2009
Published online 21 January 2010;
10.1126/science.1182378
Include this information when citing this paper.

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

Sharon R. Grossman,[1,2*]† Ilya Shylakhter,[1,2*]† Elinor K. Karlsson,[1,2] Elizabeth H. Byrne,[1,2] Shannon Morales,[1,2,3] Gabriel Frieden,[1] Elizabeth Hostetter,[1,2] Elaine Angelino,[1,4] Manuel Garber,[2] Or Zuk,[2] Eric S. Lander,[2,4,5] Stephen F. Schaffner,[2] Pardis C. Sabeti[1,2,4]†

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. By applying CMS to candidate regions from the International Haplotype Map, we localized population-specific selective signals to 55 kilobases (median), identifying known and novel causal variants. CMS can not just identify 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 (hundreds 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.

[1]Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [3]Mount Sinai School of Medicine, New York, NY 10029, USA. [4]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. [5]Department of Biology, MIT, Cambridge, MA 02139, 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.)

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: (i) 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. (ii) High-frequency derived alleles: A new (nonancestral, or derived) allele rises to a frequency higher than expected under genetic drift, carrying neighboring derived alleles with it. (iii) 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 30,000 to 80,000 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) (*3*, *4*) and one for population differentiation (F_{ST}) (*5*). 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 to 30,000 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 1 Mb of simulated sequence data (~10,000 polymorphisms) for 120 chromosomes from each population. In addition, we generated a data set 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 F_{ST} 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 and fig. S2). In contrast, iHS, ΔiHH, 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 nonselected variants and were only weakly correlated for neutral variants, we combined them in a composite likelihood statistic, termed the composite of multiple signals

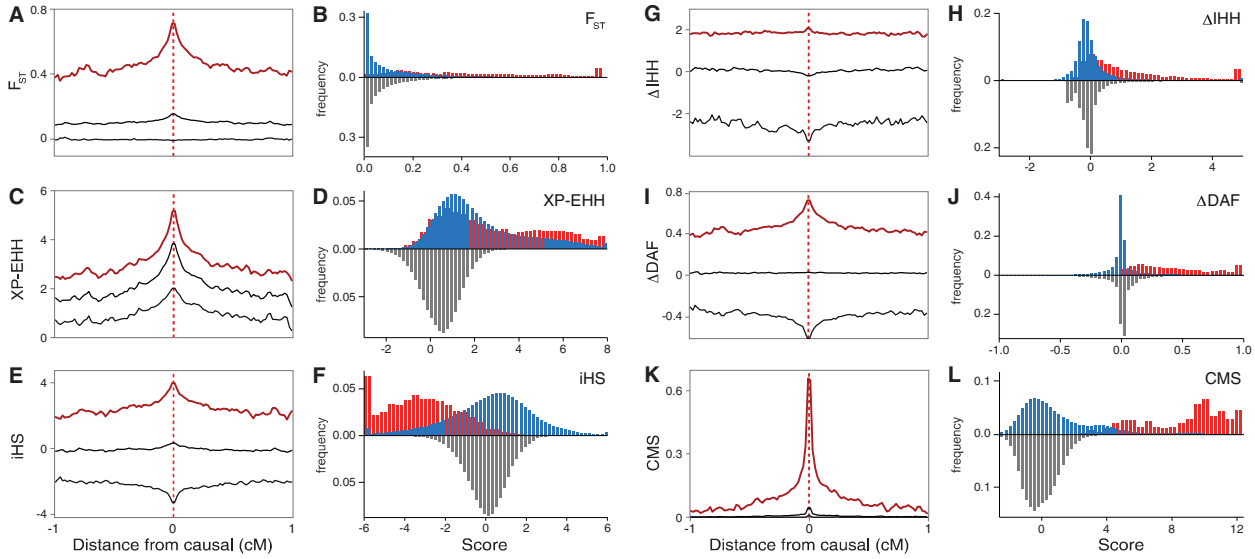

Fig. 1. CMS localizes selection and identifies causal variants better than single tests. (**A**, **C**, **E**, **G**, **I**) Top 5% (red line) and bottom 5% (black line) of scores and mean score (black, dashed line) in 1 MB surrounding causal mutation (located at red dashed line). (**B**, **D**, **F**, **H**, **J**) Distribution of scores for the causal variant (red bars), nearby unselected variants (blue bars), and variants in regions without selection (gray bars, below axis). The composite test (CMS) outperforms individual tests for (**K**) localizing the selective signal and (**L**) distinguishing the causal variant.

(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 π, the CMS score is the approximate posterior probability that the variant is selected:

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

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 [supporting online material (SOM)].

In simulations, CMS showed power both to localize the selection signal spatially and distinguish the causal variant (Fig. 1, K and L). Whereas single tests provided weak localization (~1 Mb), CMS localized the signal to an average 89 kb (for full sequence data) and contained the causal variant in 90% of cases. With sparser genotype data (corresponding to HapMapII), CMS localized to 104 kb, even when the causal variant was absent from the data set. 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 to 1500 candidate causal variants per region, whereas CMS narrowed the signal to ~100 (table S2). The causal variant was among the top 20 variants in half of cases and was the highest-scoring variant in a quarter of cases, with high power given that we included sweeps to frequencies as low as 20%. The power for

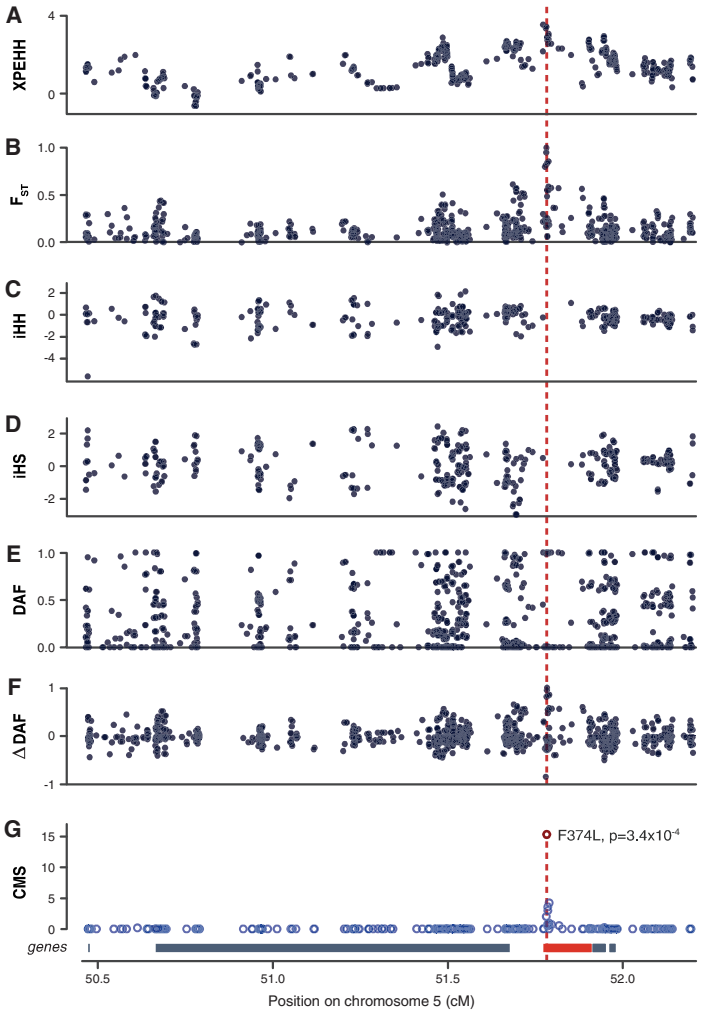

Fig. 2. Localizing selection at *MATP*. Scores of six individual tests (**A** to **F**) and CMS (**G**) for a region containing *MATP*. A nonsynonymous SNP [rs16891982, F374L (Phe^{374}→Leu), red dotted line] associated with pigmentation is believed to be the mutation under selection.

sweeps where the causal allele is at high frequency (>50%) is even greater, with the causal variant among the top 10 variants in half of cases (table S3).

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)

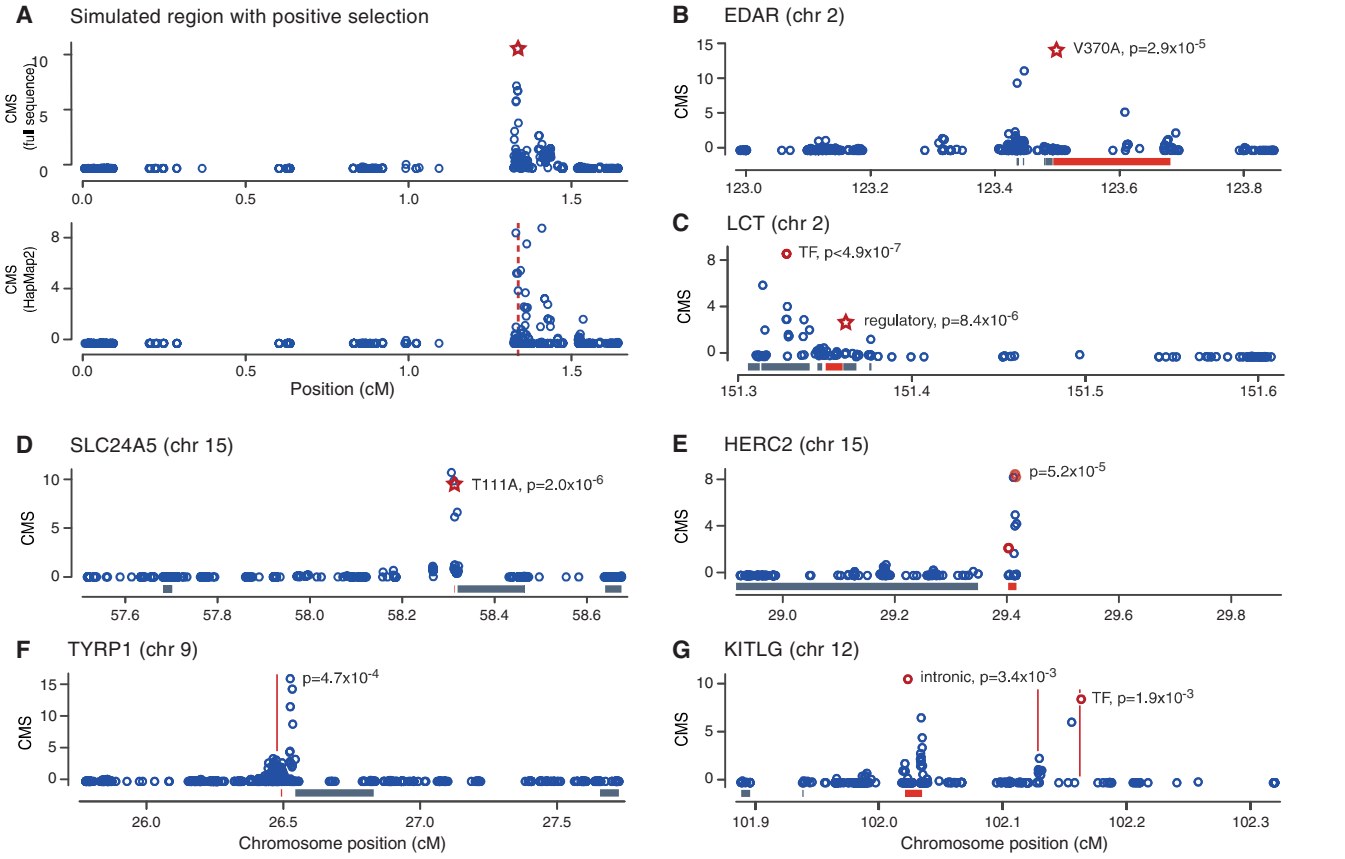

Fig. 3. CMS localizes selection and identifies causal variants in simulated and empirical data. CMS analysis of (**A**) simulated full sequence and HapMapII-density genotype data sets; and HapMapII selective sweeps at the genes (**B**) *EDAR*, (**C**) *LCT*, (**D**) *SLC24A5*, (**E**) *OCA2/HERC2*, (**F**) *TYRP1*, and (**G**) *KITLG*. Bars on *x* axis indicate genes (red bars: putative selected gene; gray bars: other genes); blue circles show CMS values; red stars indicate putative causal alleles; red circles indicate SNPs with annotated function and/or trait association; red lines mark other associated loci.

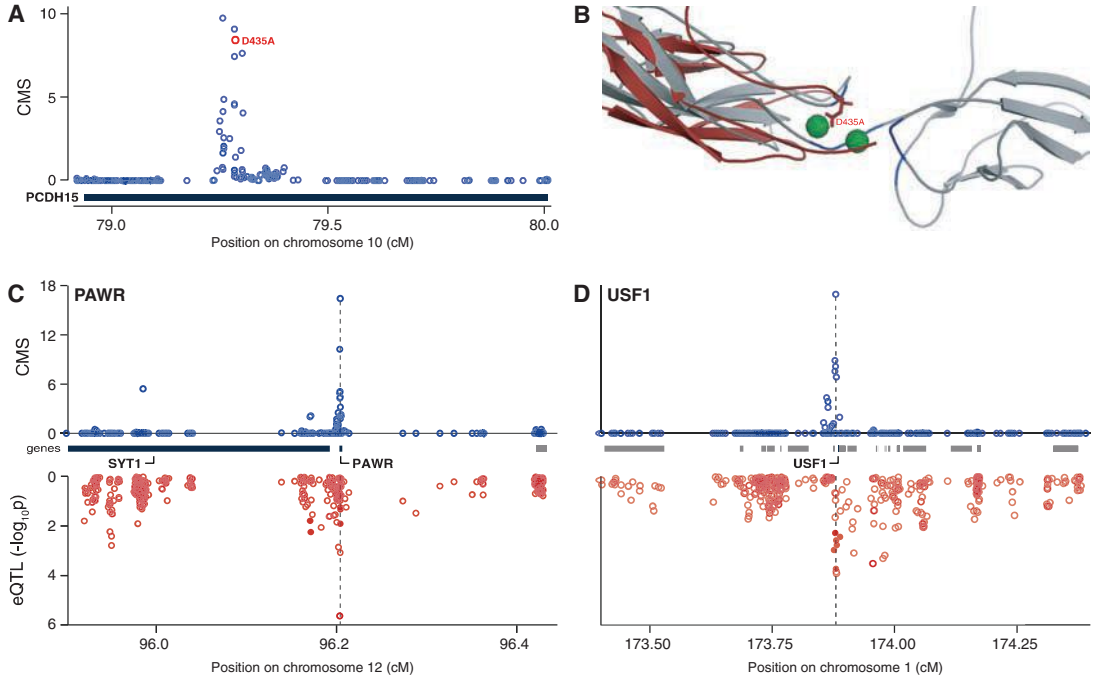

Fig. 4. Coding and regulatory mutations identified by CMS. (**A**) CMS scores around *PCDH15* (HapMapII data). Red circle: nonsynonymous mutation (D435A). (**B**) Homology modeling of the PCDH15 cadherin-4 domain (red) predicts that D435A (red rods) is among the residues (blue) coordinating calcium ions (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 the 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).

(*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 data set 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, *SLC24A5*, *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 10 CMS scores, out of 1000 to 1500 variants in the region. We also examined four regions (350 kb to 1 MB) 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- to 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 and 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^{2+}-binding site at the interface of cadherin repeats in the protein's extracellular domain (SOM) (Fig. 4A and 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 single-nucleotide polymorphisms (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 about eight to about one, the method reveals instances of multiple genes in the same pathway showing signs of selection. For example, in addition to *PCDH15*, four genes linked to cochlear function or Usher syndrome (*1*, *16*) show evidence for selection in East Asia. We used the PANTHER Gene Ontology database to test for this enrichment on all CMS-localized regions from HapMapII (SOM) (*17*). We found statistically significant enrichment for several categories (table S7): sensory perception genes (including *PCDH15*) are enriched for selection in East Asia, immune-related genes in West Africa, and genes related to homeostasis and metabolism in all three populations.

CMS can narrow candidate regions for recent local adaptation in humans and identify small numbers of candidate polymorphisms. For this kind of event, we may already be close to the limit on localization from population signals alone. According to our simulations, each causal variant has on average 20 perfect proxies (fig. S10), all essentially indistinguishable from the causal variant. Identifying specific causal variants may thus require functional characterization of small sets of candidates.

The CMS method can be adapted to a wider range of selective regimes, including detecting (i) older selection occurring any time after the divergence of human populations (50,000 to 75,000 years) (F_{ST} and ΔDAF would here become the predominant CMS signals) and (ii) selection on standing variation or very old selection (by incorporating additional population-based tests). It can be applied to nonhuman species with population samples of dense genotype or sequence data; as these increasingly become available, the details of the appropriate CMS test would depend on the demographic history and population structure of the species.

Within human genetics, the research community is currently generating data sets of human variation in many populations, through initiatives such as the 1000 Genomes Project (*18*). With continuing improvements in sequencing technology, 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

1. J. M. Akey, *Genome Res.* **19**, 711 (2009).
2. P. C. Sabeti *et al.*, *Science* **312**, 1614 (2006).
3. P. C. Sabeti *et al.*; International HapMap Consortium, *Nature* **449**, 913 (2007).
4. B. F. Voight, S. Kudaravalli, X. Wen, J. K. Pritchard, *PLoS Biol.* **4**, e72 (2006).
5. B. S. Weir, C. C. Cockerham, *Evolution* **38**, 1358 (1984).
6. S. F. Schaffner *et al.*, *Genome Res.* **15**, 1576 (2005).
7. K. A. Frazer *et al.*; International HapMap Consortium, *Nature* **449**, 851 (2007).
8. K. M. Teshima, G. Coop, M. Przeworski, *Genome Res.* **16**, 702 (2006).
9. R. A. Sturm, *Hum. Mol. Genet.* **18**, (R1), R9 (2009).
10. S. H. Williamson *et al.*, *PLoS Genet.* **3**, e90 (2007).
11. L. Shapiro *et al.*, *Nature* **374**, 327 (1995).
12. Z. M. Ahmed *et al.*, *Hum. Mol. Genet.* **12**, 3215 (2003).
13. P. Kazmierczak *et al.*, *Nature* **449**, 87 (2007).
14. K. S. Park *et al.*, *J. Hum. Genet.* **51**, 85 (2006).
15. C. Y. Lin *et al.*, *PLoS Genet.* **3**, e87 (2007).
16. J. Reiners *et al.*, *Hum. Mol. Genet.* **14**, 3933 (2005).
17. P. D. Thomas *et al.*, *Nucleic Acids Res.* **31**, 334 (2003).
18. www.1000genomes.org

19. We thank K. Andersen, J. Lohmueller, B. Stranger, S. McCarroll, K. Lohmueller, K. Lindblad-Toh, M. Guttman, and J. Rinn for functional guidance, and E. Phelan, D. Altshuler, and the Sabeti Lab for helpful discussions throughout. P.C.S. is supported by the Burroughs Wellcome and Packard foundations. E.K.K. 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

27 October 2009; accepted 16 December 2009
Published online 7 January 2010;
10.1126/science.1183863
Include this information when citing this paper.

ERRATUM

Post date 17 February 2012

Reports: "A composite of multiple signals distinguishes causal variants in regions of positive selection" by S. R. Grossman *et al.* (12 February 2010, p. 883). The surname of the second author was spelled incorrectly; the correct spelling is Shlyakhter. The name has been corrected in the HTML version online.