pression was undetectable. There has been some disagreement about the timing of expression of *Myf5* and *MyoD* in the branchial arches, depending on the method of detection, but the earliest reported expression of these genes in this region is E9.25 and E9.5, respectively (*22–24*). By E9.5, *Myf5* and *capsulin* were expressed in the same cell population within the first branchial arch, and by E10.5, *Myf5*, *capsulin*, and *MyoR* were coexpressed in these cells of wild-type embryos (Fig. 3A). In contrast, *Myf5* was not expressed in first branchial arch precursors of $MyoR^{-/-}capsulin^{-/-}$ double mutants at E9.5 or E11.5 (Fig. 3B). There was also no evidence for expression of *Myf5*, *MyoD*, or *myogenin* at E15.5 in the region of affected facial muscles (Fig. 3C), whereas these genes were expressed in other developing head and trunk muscles.

To determine the fate of first arch muscle precursors that failed to activate expression of *Myf5* and *MyoD*, we performed TUNEL (terminal deoxynucleotidyl transferase-mediated dUTP nick-end labeling) on histological sections of double-mutant embryos at E10.5, when cells marked by expression of *capsulin-lacZ* were disappearing. As shown in Fig. 4, TUNEL-positive cells were observed among the lacZ-positive muscle precursors of double mutants, but not of $MyoR^{-/-}capsulin^{+/-}$ embryos. We conclude that these cells, which fail to initiate the normal program for muscle development in the double mutant, undergo apoptosis with resulting ablation of muscles of mastication. Similar observations have been made in muscle precursor cells in the limb buds of mice lacking *MyoD* and *myf5* (*25*).

The absence of specific head muscle cells, as well as markers of the corresponding myogenic lineages, in $MyoR^{-/-}capsulin^{-/-}$ mutants resembles the effect of $MyoD^{-/-}Myf5^{-/-}$ double mutations on all skeletal muscles (*2*) and is distinct from the phenotype of $Myf5^{-/-}$ $Pax3^{-/-}$ mutants, which exhibit a specific deficiency of trunk skeletal muscles (*7*). This phenotype also differs from that of *myogenin* mutant mice, in which myoblasts express myogenic bHLH genes, but are unable to differentiate (*3*, *4*). These findings demonstrate that *MyoR* and *capsulin* redundantly regulate an initial step in the specification of a specific subset of facial skeletal muscle lineages and that, in the absence of these factors, myogenic bHLH genes are not switched on, and cells from these lineages undergo programmed cell death. There may also be a modest effect on migration of precursors, as is seen in *Lbx1* mutant mice (*21*). MyoR and capsulin act as transcriptional repressors in transfection assays (*12*, *20*). Whether they act to repress an inhibitor of myogenesis or have a transcriptional-activating function during development of facial muscle remains to be determined. The phenotype of $MyoR^{-/-}capsulin^{-/-}$ mutant mice reveals a previously unanticipated complexity in the development of head skeletal muscles, and these findings identify MyoR and capsulin as unique transcriptional regulators for the development of specific head muscles.

References and Notes

1. M. Buckingham. *Curr. Opin. Genet. Dev.* **11**, 440 (2001).
2. M. A. Rudnicki *et al.*, *Cell* **75**, 1351 (1993).
3. P. Hasty *et al.*, *Nature* **364**, 501 (1993).
4. Y. Nabeshima *et al.*, *Nature* **364**, 532 (1993).
5. A. Rawls, M. R. Valdez, W. Zhang, J. Richardson, W. H. Klein, E. N. Olson. *Development* **125**, 2349 (1998).
6. A. Rawls, E. N. Olson, *Cell* **89**, 5 (1997).
7. S. Tajbakhsh, D. Rocancourt, G. Cossu, M. Buckingham, *Cell* **89**, 127 (1997).
8. B. Christ, C. P. Ordahl, *Anat. Embryol. (Berlin)* **191**, 381 (1995).
9. D. M. Noden, *Am. J. Anat.* **168**, 257 (1983).
10. P. A. Trainor, S. S. Tan, P. P. Tam, *Development* **120**, 2397 (1994).
11. A. Hacker, S. Guthrie, *Development* **125**, 3461 (1998).
12. J. Lu, R. Webb, J. A. Richardson, E. N. Olson, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 552 (1999).
13. L. Robb, L. Hartley, C. C. Wang, R. P. Harvey, C. G. Begley, *Mech. Dev.* **76**, 197 (1998).
14. J. Lu, J. A. Richardson, E. N. Olson, *Mech. Dev.* **73**, 23 (1998).
15. H. Hidai, R. Bardales, R. Goodwin, T. Quertermous, E. E. Quertermous, *Mech. Dev.* **73**, 33 (1998).
16. S. E. Quaggin *et al.*, *Development* **126**, 5771 (1999).
17. L. Robb *et al.*, *Dev. Dyn.* **213**, 105 (1998).
18. J. Lu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 9525 (2000).
19. S. E. Quaggin *et al.*, *Development* **126**, 5771 (1999).
20. J. Lu, E. N. Olson, unpublished results.
21. H. Brohmann, K. Jagla, C. Birchmeier, *Development* **127,** 437 (2000).
22. M. O. Ott, E. Bober, G. Lyons, H. Arnold, M. Buckingham, *Development* **111**, 1097 (1991).
23. S. Tajbakhsh, E. Bober, C. Babinet, S. Pournin, H. Arnold, M. Buckingham, *Dev. Dyn.* **206,** 291 (1996).
24. J. C. J. Chen, C. M. Love, D. J. Goldhammer. *Dev. Dyn.* **221,** 274 (2001).
25. B. Kablar, K. Krastel, C. Ying, S. J. Tapscott, D. J. Goldhammer, M. A. Rudnicki. *Dev. Biol.* **206**, 21931 (1999).

26. We are grateful to C. Pomajzl and J. Stark for histologic preparations. We also thank A. Tizenor for graphics and J. Page for editorial assistance. Supported by grants from the NIH, the Donald W. Reynolds Foundation and the Muscular Dystrophy Association to E.N.O.

Supporting Online Material
www.sciencemag.org/cgi/content/full/298/5602/2378/DC1
Materials and Methods
Fig. S1
References

9 September 2002; accepted 28 October 2002

Genetic Structure of Human Populations

Noah A. Rosenberg,[1]* Jonathan K. Pritchard,[2] James L. Weber,[3] Howard M. Cann,[4] Kenneth K. Kidd,[5] Lev A. Zhivotovsky,[6] Marcus W. Feldman[7]

We studied human population structure using genotypes at 377 autosomal microsatellite loci in 1056 individuals from 52 populations. Within-population differences among individuals account for 93 to 95% of genetic variation; differences among major groups constitute only 3 to 5%. Nevertheless, without using prior information about the origins of individuals, we identified six main genetic clusters, five of which correspond to major geographic regions, and subclusters that often correspond to individual populations. General agreement of genetic and predefined populations suggests that self-reported ancestry can facilitate assessments of epidemiological risks but does not obviate the need to use genetic information in genetic association studies.

Most studies of human variation begin by sampling from predefined "populations." These populations are usually defined on the basis of culture or geography and might not reflect underlying genetic relationships (*1*). Because knowledge about genetic structure of modern human populations can aid in inference of human evolutionary history, we used the HGDP-CEPH Human Genome Diversity Cell Line Panel (*2, 3*) to test the correspondence of predefined groups with those inferred from individual multilocus genotypes (supporting online text).

The average proportion of genetic differences between individuals from different human populations only slightly exceeds that between unrelated individuals from a single population (*4–9*). That is, the within-population component of genetic variation, estimated here as 93 to 95% (Table 1), accounts for most of human genetic diversity. Perhaps as a result of differences in sampling schemes (*10*), our estimate is higher than previous estimates from studies of comparable geographic coverage (*4–6*, *9*), one of which also used microsatellite markers (*6*). This overall similarity of human populations is also evident in the geographically widespread nature of most alleles (fig. S1). Of 4199 alleles present more than once in the sample, 46.7% appeared in all major regions represented: Africa, Europe, the Middle East, Central/

South Asia, East Asia, Oceania, and America. Only 7.4% of these 4199 alleles were exclusive to one region; region-specific alleles were usually rare, with a median relative frequency of 1.0% in their region of occurrence (*11*).

Despite small among-population variance components and the rarity of "private" alleles, analysis of multilocus genotypes allows inference of genetic ancestry without relying on information about sampling locations of individuals (*12–14*). We applied a model-based clustering algorithm that, loosely speaking, identifies subgroups that have distinctive allele frequencies. This procedure, implemented in the computer program *structure* (*14*), places individuals into K clusters, where K is chosen in advance but can be varied across independent runs of the algorithm. Individuals can have membership in multiple clusters, with membership coefficients summing to 1 across clusters.

In the worldwide sample, individuals from the same predefined population nearly always shared similar membership coefficients in inferred clusters (Fig. 1). At $K = 2$ the clusters were anchored by Africa and America, regions separated by a relatively large genetic distance (table S1). Each increase in K split one of the clusters obtained with the previous value. At $K = 5$, clusters corresponded largely to major geographic regions. However, the next cluster at $K = 6$ did not match a major region but consisted largely of individuals of the isolated Kalash group, who speak an Indo-European language and live in northwest Pakistan (Fig. 1 and table S2). In several populations, individuals had partial membership in multiple clusters, with similar membership coefficients for most individuals. These populations might reflect continuous gradations in allele frequencies across regions or admixture of neighboring groups. Unlike other populations from Pakistan, Kalash showed no membership in East Asia at $K = 5$, consistent with their suggested European or Middle Eastern origin (*15*).

In America and Oceania, regions with low heterozygosity (table S3), inferred clusters corresponded closely to predefined populations (Fig. 2). These regions had the largest among-population variance components, and they required the fewest loci to obtain the clusters observed with the full data. Inferred clusters for Africa and the Middle East were also consistent across runs but did not all correspond to predefined groups. For the other samples, among-population variance components were below 2%, and independent *structure* runs were less consistent. For $K \geq 3$, similarity coefficients for pairs of runs

[1]Molecular and Computational Biology, 1042 West 36th Place DRB 289, University of Southern California, Los Angeles, CA 90089, USA. [2]Department of Human Genetics, University of Chicago, 920 East 58th Street, Chicago, IL 60637, USA. [3]Center for Medical Genetics, Marshfield Medical Research Foundation, Marshfield, WI 54449, USA. [4]Foundation Jean Dausset–Centre d'Etude du Polymorphisme Humain (CEPH), 27 rue Juliette Dodu, 75010 Paris, France. [5]Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06520, USA. [6]Vavilov Institute of General Genetics, Russian Academy of Sciences, 3 Gubkin Street, Moscow 117809, Russia. [7]Department of Biological Sciences, Stanford University, Stanford, CA 94305, USA.

*To whom correspondence should be addressed. E-mail: noahr@usc.edu

Table 1. Analysis of molecular variance (AMOVA). Eurasia, which encompasses Europe, the Middle East, and Central/South Asia, is treated as one region in the five-region AMOVA but is subdivided in the seven-region design. The World-B97 sample mimics a previous study (*6*).

Sample	Number of regions	Number of populations	Variance components and 95% confidence intervals (%)		
			Within populations	Among populations within regions	Among regions
World	1	52	94.6 (94.3, 94.8)	5.4 (5.2, 5.7)	
World	5	52	93.2 (92.9, 93.5)	2.5 (2.4, 2.6)	4.3 (4.0, 4.7)
World	7	52	94.1 (93.8, 94.3)	2.4 (2.3, 2.5)	3.6 (3.3, 3.9)
World-B97	5	14	89.8 (89.3, 90.2)	5.0 (4.8, 5.3)	5.2 (4.7, 5.7)
Africa	1	6	96.9 (96.7, 97.1)	3.1 (2.9, 3.3)	
Eurasia	1	21	98.5 (98.4, 98.6)	1.5 (1.4, 1.6)	
Eurasia	3	21	98.3 (98.2, 98.4)	1.2 (1.1, 1.3)	0.5 (0.4, 0.6)
Europe	1	8	99.3 (99.1, 99.4)	0.7 (0.6, 0.9)	
Middle East	1	4	98.7 (98.6, 98.8)	1.3 (1.2, 1.4)	
Central/South Asia	1	9	98.6 (98.5, 98.8)	1.4 (1.2, 1.5)	
East Asia	1	18	98.7 (98.6, 98.9)	1.3 (1.1, 1.4)	
Oceania	1	2	93.6 (92.8, 94.3)	6.4 (5.7, 7.2)	
America	1	5	88.4 (87.7, 89.0)	11.6 (11.0, 12.3)	

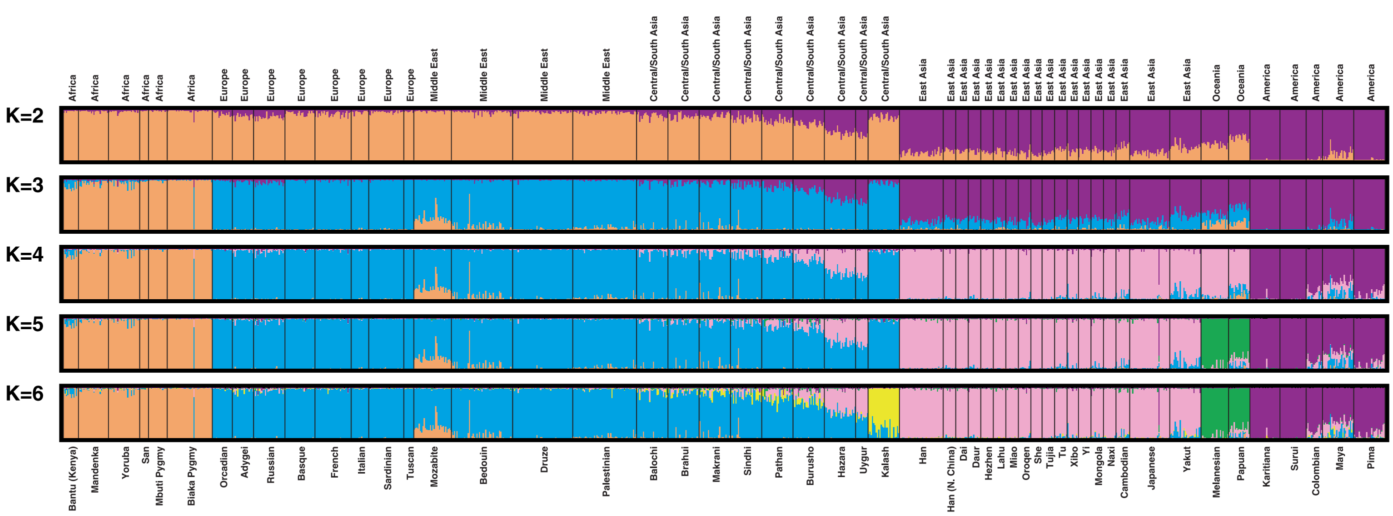

Fig. 1. Estimated population structure. Each individual is represented by a thin vertical line, which is partitioned into K colored segments that represent the individual's estimated membership fractions in K clusters. Black lines separate individuals of different populations. Populations are labeled below the figure, with their regional affiliations above it. Ten *structure* runs at each K produced nearly identical individual membership coefficients, having pairwise similarity coefficients above 0.97, with the exceptions of comparisons involving four runs at $K = 3$ that separated East Asia instead of Eurasia, and one run at $K = 6$ that separated Karitiana instead of Kalash. The figure shown for a given K is based on the highest probability run at that K.

were typically moderate (0.1 to 0.85), rather than large (0.85 to 1.0). However, various patterns were observed across runs.

In East Asia, Yakut, whose language is Altaic, and Japanese, whose language is often classified as Altaic, were usually identified as distinctive. Other speakers of Altaic languages, including Daur, Hezhen, Mongola, Oroqen, and Xibo, all from northern China, shared a greater degree of membership with Japanese and Yakut than with more southerly groups from other language families, such as Cambodian, Dai, Han, Miao, Naxi, She, Tujia, and Yi. However, Tu, who speak an Altaic language and live in north-central China, largely grouped with the southern populations. Lahu, who speak a Sino-Tibetan language and were the least heterozygous population in the region, frequently separated despite their proximity with other groups sampled from southern China (*16*).

Eurasia frequently separated into its component regions, along with Kalash. Adygei, from the Caucasus, shared membership in Europe and Central/South Asia. Within Central/South Asia, Burusho of northern Pakistan, a linguistic isolate, largely separated from other groups, although less clearly than the genetic isolate, Kalash. Perhaps as a result of shared Mongol ancestry (*15, 16*), Hazara of Pakistan and Uygur of northwestern China, whose languages are Indo-European and Altaic, respectively, clustered together. For Balochi, Makrani, Pathan, and Sindhi, all of whose languages are Indo-European, and less so for Dravidian-speaking Brahui, multiple clusters were found, with individuals from many populations having membership in each cluster.

Europe, with the smallest among-population variance component (0.7%), was the most difficult region in which to detect population structure. The highest-likelihood run for $K = 3$ found no structure; in other runs, Basque and Sardinian were identified as distinctive. Russians variously grouped with Adygei and Orcadians; Russian-Orcadian similarity might derive from shared Viking contributions (*17*). French, Italians, and Tuscans showed mixed membership in clusters that contained other populations.

Because genetic drift occurs rapidly in small populations, particularly in those that are also isolated, these groups quickly accu-

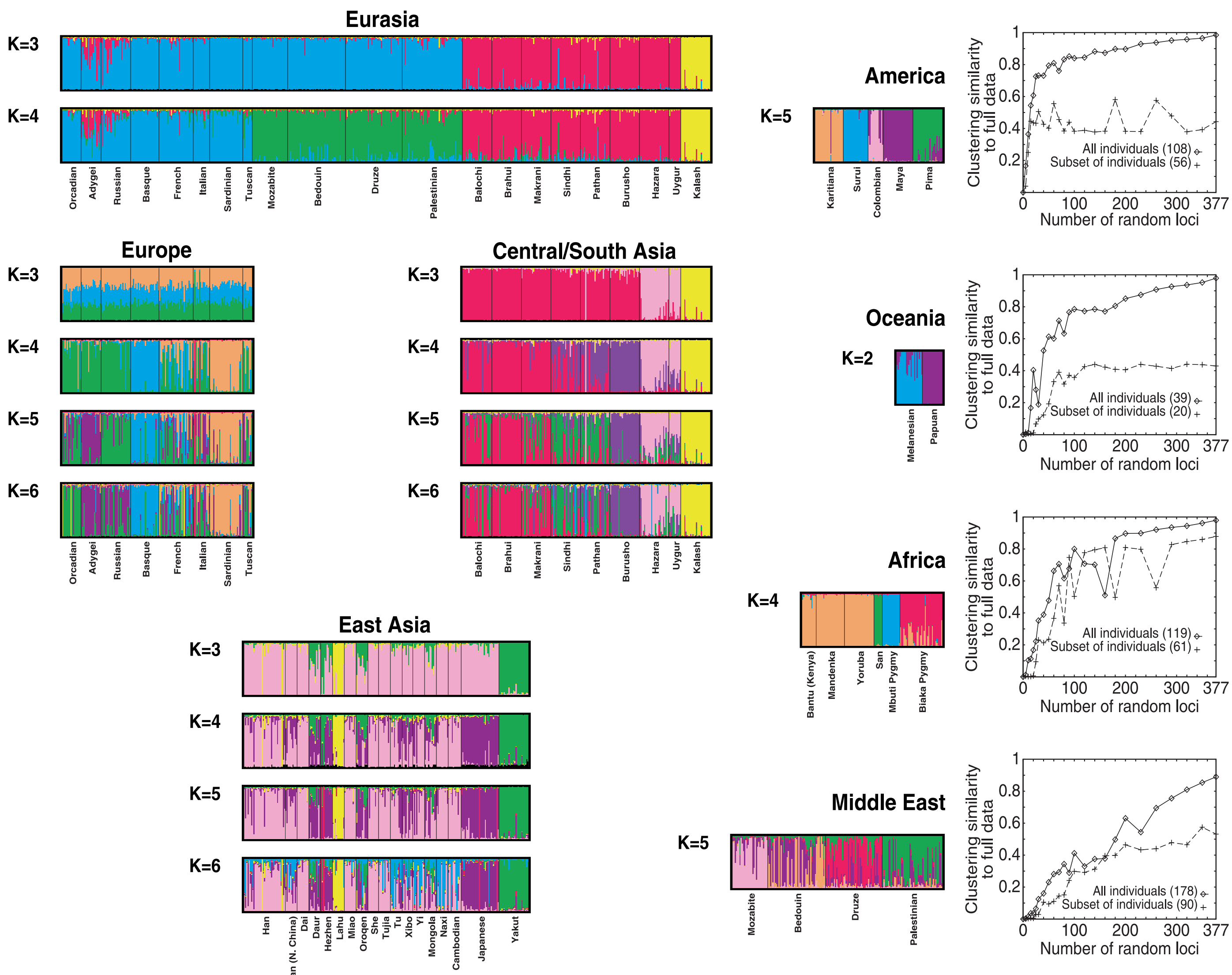

Fig. 2. Estimated population structure for regions. For America, Oceania, Africa, and the Middle East, solutions were consistent across 10 runs (all similarity coefficients above 0.97, 0.93, 0.97, and 0.86, respectively, except those involving one run with Africa that assigned many Biaka individuals partial membership with San). Values of *K* shown for these samples are the highest values for which this was true, and the highest probability runs are shown. For remaining regions, solutions were more variable across runs, and the highest probability runs for various values of *K* are displayed. Graphs for America, Oceania, Africa, and the Middle East display median similarity coefficients between runs based on the full data and runs based on subsets of the data. Correspondence of colors across figures for different regions is not meaningful.

mulate distinctive allele frequencies. Thus, *structure* efficiently detects isolated and relatively homogeneous groups, even if the times since their divergences or exchanges with other groups are short (*18*). This phenomenon may explain the inferred distinctiveness of groups with low heterozygosity, such as Lahu and American groups, and those that are small and isolated, such as Kalash. Groups with larger sample sizes are also more easily separated; thus, the difficulty of clustering in East Asia was exacerbated by small sample sizes. Because sampling was population based, the sample likely produced clusters that were more distinct than would have been found in a sample with random worldwide representation. However, world-level boundaries between major clusters mostly corresponded to major physical barriers (oceans, Himalayas, Sahara).

The amount of among-group variation affects the number of loci required to produce clusters similar to those obtained with the full data. For the Middle East, with an among-population variance component of 1.3%, nearly all the loci were required to achieve a similarity of 0.8 to the clustering on the basis of full data, and use of more loci would likely produce more consistent clustering. For Oceania and Africa, only ~200 loci were needed; for the world sample, ~150 were needed (fig. S2), and ~100 were sufficient for America. Fewer loci would probably suffice for larger samples (*18*); conversely, accuracy decreased considerably when only half the sample was used (Fig. 2). The number of loci required would also decrease if extremely informative markers, such as those with particularly high heterozygosity (table S4), were genotyped (*18*). The loci here form a panel intended for use primarily in individuals of European descent (*19*). Although 10 of the loci had heterozygosity less than 0.5 in East Asia, none had similarly low European heterozygosities; thus, inference of subclusters using "random" markers might be more difficult than observed here, especially in Europe. However, the effect of excluding markers with low European heterozygosity is likely minimal, because generally high microsatellite heterozygosities ensure that relatively few loci are discarded on these grounds (*20*). The fact that regional heterozygosities here (table S3) follow the same relative order as and have nearly equal values to those of loci that were ascertained in a geographically diverse panel (*12*) provides further evidence that the ascertainment effect on heterozygosity estimates and on statistics derived from these estimates, such as genetic variance components (*21*), is small.

Genetic clusters often corresponded closely to predefined regional or population groups or to collections of geographically and linguistically similar populations. Among exceptions, linguistic similarity did not provide a general explanation for genetic groupings of populations that were relatively distant geographically, such as Hazara and Uygur or Tu and populations from southern China. Our finer clustering results compared with other multilocus studies derive from our use of more data. General correspondence between regional affiliation and genetic ancestry has been reported (*12–14*), with clearer correspondence in studies that used more loci (*13*) than in those that used fewer loci (*9*, *22*); we have further identified correspondence between genetic structure and population affiliation in regions with among-population variance components larger than 2 to 3%.

The structure of human populations is relevant in various epidemiological contexts. As a result of variation in frequencies of both genetic and nongenetic risk factors, rates of disease and of such phenotypes as adverse drug response vary across populations (*22*, *23*). Further, information about a patient's population of origin might provide health-care practitioners with information about risk when direct causes of disease are unknown (*23*). Recent articles have considered whether it is preferable to use self-reported population ancestry or genetically inferred ancestry in such situations (*22–25*). We have found that predefined labels were highly informative about membership in genetic clusters, even for intermediate populations, in which most individuals had similar membership coefficients across clusters. Sizable variation in ancestry within predefined populations was detected only rarely, such as among geographically proximate Middle Eastern groups.

Thus, for many applications in epidemiology, as well as for assessing individual disease risks, self-reported population ancestry likely provides a suitable proxy for genetic ancestry. Self-reported ancestry can be obtained less intrusively than genetic ancestry, and if self-reported ancestry subdivides a genetic cluster into multiple groups, it may provide useful information about unknown environmental risk factors (*23*, *25*). One exception to these general comments may arise in recently admixed populations, in which genetic ancestry varies substantially among individuals; this variation might correlate with risk as a result of genetic or cultural factors (*24*). In some contexts, however, use of genetic clusters is more appropriate than use of self-reported ancestry. In genetic case-control association studies, false positives can be obtained if disease risk is correlated with genetic ancestry (*24*, *26*). Basing analyses on self-reported ancestry reduces the proportion of false positives considerably (*25*). However, association studies are usually analyzed by significance testing, in which slight differences in genetic ancestry between cases and controls can produce statistically significant false-positive associations in large samples. Thus, errors incurred by using self-reported rather than genetic ancestry might cause serious problems in large studies that will be required for identifying susceptibility loci with small effects (*26*). Genetic clustering is also more appropriate for some types of population genetic studies, because unrecognized genetic structure can produce false positives in statistical tests for population growth or natural selection (*27*).

The challenge of genetic studies of human history is to use the small amount of genetic differentiation among populations to infer the history of human migrations. Because most alleles are widespread, genetic differences among human populations derive mainly from gradations in allele frequencies rather than from distinctive "diagnostic" genotypes. Indeed, it was only in the accumulation of small allele-frequency differences across many loci that population structure was identified. Patterns of modern human population structure discussed here can be used to guide construction of historical models of migration and admixture that will be useful in inferential studies of human genetic history.

References and Notes

1. M. W. Foster, R. R. Sharp, *Genome Res.* **12**, 844 (2002).
2. H. M. Cann *et al.*, *Science* **296**, 261 (2002).
3. Genotypes from this study are available at http://research.marshfieldclinic.org/genetics/Freq/FreqInfo.htm.
4. R. C. Lewontin, *Evol. Biol.* **6**, 381 (1972).
5. B. D. H. Latter, *Am. Nat.* **116**, 220 (1980).
6. G. Barbujani, A. Magagni, E. Minch, L. L Cavalli-Sforza, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 4516 (1997).
7. L. B. Jorde *et al.*, *Am. J. Hum. Genet.* **66**, 979 (2000).
8. R. A. Brown, G. J. Armelagos, *Evol. Anthropol.* **10**, 34 (2001).
9. C. Romualdi *et al.*, *Genome Res.* **12**, 602 (2002).
10. Smaller within-population variance components of comparable studies may result from their use of isolated and geographically well-separated populations to construct samples. Such a scheme might exaggerate among-group differences compared with those in the present sample, which had a smaller proportion of such populations. Indeed, when we restricted analysis to a set of populations that approximated a previous data set (*6*), we obtained a larger among-region component. Variance components also depend on sample sizes and on marker properties (*7–9*). Differential natural selection on protein variants across geographic regions might exaggerate among-group differences. Conversely, for a fixed level of within-group diversity, recurrent microsatellite mutations reduce among-group differences in comparison with those observed at markers for which each mutation produces a novel allele (*28*).
11. Recurrent mutation might be expected to influence allelic distributions considerably. However, widespread distributions of most alleles and the paucity of alleles found only in two disconnected regions suggest that recurrent mutations are only rarely followed by independent drift to sizable frequencies in multiple regions (*29*).
12. A. M. Bowcock *et al.*, *Nature* **368**, 455 (1994).
13. J. L. Mountain, L. L. Cavalli-Sforza, *Am. J. Hum. Genet.* **61**, 705 (1997).
14. J. K. Pritchard, M. Stephens, P. Donnelly, *Genetics* **155**, 945 (2000).
15. R. Qamar *et al.*, *Am. J. Hum. Genet.* **70**, 1107 (2002).
16. R. Du, V. F. Yip, *Ethnic Groups in China* (Lubrecht and Cramer, Port Jervis, NY, 1996).

17. J. Haywood, *The Penguin Historical Atlas of the Vikings* (Penguin Books, London, 1995).
18. N. A. Rosenberg *et al.*, *Genetics* **159**, 699 (2001).
19. J. L. Weber, K. W. Broman, *Adv. Genet.* **42**, 77 (2001).
20. A. R. Rogers, L. B. Jorde, *Am. J. Hum. Genet.* **58**, 1033 (1996).
21. M. Urbanek, D. Goldman, J. C. Long, *Mol. Biol. Evol.* **13**, 943 (1996).
22. J. F. Wilson *et al.*, *Nature Genet.* **29**, 265 (2001).
23. N. Risch, E. Burchard, E. Ziv, H. Tang, *Genome Biol.* **3**, comment2007.1 (2002).
24. D. C. Thomas, J. S. Witte, *Cancer Epidemiol. Biomark. Prev.* **11**, 505 (2002).
25. S. Wacholder, N. Rothman, N. Caporaso, *Cancer Epidemiol. Biomark. Prev.* **11**, 513 (2002).
26. J. K. Pritchard, P. Donnelly, *Theor. Popul. Biol.* **60**, 227 (2001).
27. S. E. Ptak, M. Przeworski, *Trends Genet.* **18**, 559 (2002).
28. L. Jin, R. Chakraborty, *Heredity* **74**, 274 (1995).
29. F. Calafell *et al.*, *Eur. J. Hum. Genet.* **6**, 38 (1998).
30. D. Altshuler, M. Cho, D. Falush, H. Innan, L. Kurina, J. Mountain, D. Nettle, M. Nordborg, M. Przeworski, N. Risch, D. Rosenberg, M. Stephens, D. Thomas, and E. Ziv provided helpful comments. The Mammalian Genotyping Service is supported by the National Heart, Lung, and Blood Institute. This work was supported by an NSF Biological Informatics Postdoctoral Fellowship (N.A.R.), a Burroughs-Wellcome Fund Hitchings Elion grant (J.K.P.), and NIH GM28428 (M.W.F.).

Supporting Online Material
www.sciencemag.org/cgi/content/full/298/5602/2381/DC1
Materials and Methods
Supporting Text
Figs. S1 and S2
Tables S1 to S4
References

19 June 2002; accepted 30 October 2002

NPAS2: A Gas-Responsive Transcription Factor

Elhadji M. Dioum,[1] Jared Rutter,[2] Jason R. Tuckerman,[1] Gonzalo Gonzalez,[1] Marie-Alda Gilles-Gonzalez,[1]* Steven L. McKnight[2]*

Neuronal PAS domain protein 2 (NPAS2) is a mammalian transcription factor that binds DNA as an obligate dimeric partner of BMAL1 and is implicated in the regulation of circadian rhythm. Here we show that both PAS domains of NPAS2 bind heme as a prosthetic group and that the heme status controls DNA binding in vitro. NPAS2-BMAL1 heterodimers, existing in either the apo (heme-free) or holo (heme-loaded) state, bound DNA avidly under favorably reducing ratios of the reduced and oxidized forms of nicotinamide adenine dinucleotide phosphate. Low micromolar concentrations of carbon monoxide inhibited the DNA binding activity of holo-NPAS2 but not that of apo-NPAS2. Upon exposure to carbon monoxide, inactive BMAL1 homodimers were formed at the expense of NPAS2-BMAL1 heterodimers. These results indicate that the heterodimerization of NPAS2, and presumably the expression of its target genes, are regulated by a gas through the heme-based sensor described here.

PAS domains are independently folding modules of ~130 amino acids that detect diverse environmental signals, including oxygen, light, voltage, redox potential, and many small aromatic molecules (*1–7*). Although these domains have modest sequence similarity, they share strikingly similar three-dimensional folds (*8–12*). Two groups of bacterial proteins—the FixL proteins of *Rhizobia* and the PDEA1 phosphodiesterases of *Acetobacter*—use heme bound within a PAS domain to sense oxygen (*13*). In FixL, binding of oxygen to the heme controls a kinase domain that phosphorylates a cognate transcription factor. In PDEA1, the heme-binding domain controls a phosphodiesterase domain that regulates the abundance of a cyclic nucleotide second messenger. A serendipitous discovery of apparent heme binding during the purification of NPAS2, a mammalian bHLH (basic helix-loop-helix)–PAS transcription factor, stimulated us to investigate whether NPAS2 might represent yet another heme-based mode of signal transduction by PAS domains.

Overexpression of a fragment of NPAS2 containing its bHLH DNA binding domain and both PAS domains in bacteria yielded amber-colored cells. The absorption spectra of liquid cultures containing those cells revealed a correlation between NPAS2 expression and heme protein absorption (Fig. 1A). Obvious peaks of absorption for the intact living cells were observed at 426 nm (Soret or gamma) and 561 nm (alpha). Upon centrifugation of a cell lysate, the bulk of overexpressed NPAS2 was recovered as an insoluble red suspension. The apoprotein resulting from solubilization of the material by denaturation and renaturation was easily reconstituted with free hemin (*14, 15*). The absorption peaks for the reconstituted proteins also occurred at 426 nm and 561 nm, with a lower extinction peak becoming detectable at 530 nm (beta) (Fig. 1B). To examine the stability and stoichiometry of the heme, we exposed this reconstituted material to a fivefold molar excess of $His^{64} \rightarrow Tyr$, $Val^{68} \rightarrow Phe$ apomyoglobin (apo-H64Y/V68F) (*16*). As indicated by the similar rates of apo-H64Y/V68F reconstitution with heme abstracted from either NPAS2 or *Bradyrhizobium japonicum* FixL, the heme stability of the two proteins was comparable (Fig. 1C). The final absorption value for the apo-H64Y/V68F treated with NPAS2 showed that NPAS2 had rough-

[1]Departments of Biochemistry and Plant Biology and Plant Biotechnology Center, The Ohio State University, 1060 Carmack Road, Columbus, OH 43210, USA. [2]Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390, USA.

*To whom correspondence should be addressed. E-mail: steven.mcknight@utsouthwestern.edu, magg@biochem.swmed.edu

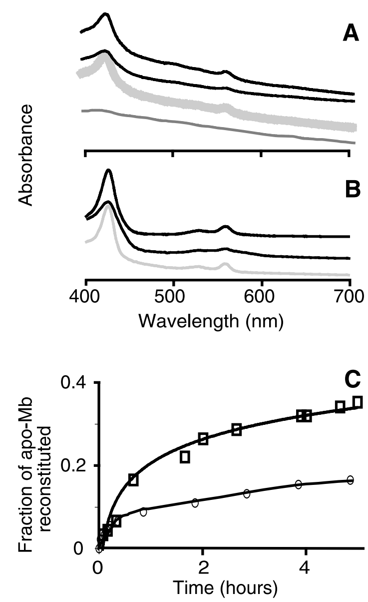

Fig. 1. Heme content and stability of holo-NPAS2. (**A**) Production of heme protein in whole *E. coli* cells after the induction of TG1 cells not expressing recombinant genes (thin gray line) or expressing the NPAS2 truncated recombinant forms bHLH–PAS-A–PAS-B (thick black line), bHLH–PAS-A (thin black line), or PAS-B (thick gray line). NPAS2 fragments, placed downstream from a tac promoter in a pUC19-derived expression vector, were expressed after 5 hours of isopropyl-β-D-thiogalactopyranoside induction in *E. coli* strain TG1. The absorption spectra of 10-fold concentrated cultures of intact cells were collected with an ATI Unicam UV-4 UV/Vis spectrophotometer (Spectronic Instruments Inc., Rochester, NY) containing a turbid-sample accessory. (**B**) Absorption spectra of the deoxy (Fe^{II}) forms of purified bHLH–PAS-A–PAS-B (thick black line), bHLH–PAS-A (thin black line), and PAS-B (gray line). Deoxy species were prepared by reducing the protein with dilute dithionite in an anaerobic glove box and rapidly transferring it, by gel filtration, to 0.10 M sodium phosphate (pH 7.5) and 5 mM dithiothreitol (DTT). (**C**) Extraction of heme from reconstituted holo-bHLH–PAS-A–PAS-B (squares) or from *B. japonicum* FixL protein (circles) by a fivefold molar excess of apo-H64Y/V68F sperm whale myoglobin at pH 6.5 and 25°C (*16*).