nature genetics

Geographical genomics of human leukocyte gene expression variation in southern Morocco

Youssef Idaghdour[1], Wendy Czika[2], Kevin V Shianna[3], Sang H Lee[4], Peter M Visscher[4], Hilary C Martin[5], Kelci Miclaus[2], Sami J Jadallah[6], David B Goldstein[3], Russell D Wolfinger[2] & Greg Gibson[1,5]

Studies of the genetics of gene expression can identify expression SNPs (eSNPs) that explain variation in transcript abundance. Here we address the robustness of eSNP associations to environmental geography and population structure in a comparison of 194 Arab and Amazigh individuals from a city and two villages in southern Morocco. Gene expression differed between pairs of locations for up to a third of all transcripts, with notable enrichment of transcripts involved in ribosomal biosynthesis and oxidative phosphorylation. Robust associations were observed in the leukocyte samples: *cis* eSNPs ($P < 10^{-08}$) were identified for 346 genes, and *trans* eSNPs ($P < 10^{-11}$) for 10 genes. All of these associations were consistent both across the three sample locations and after controlling for ancestry and relatedness. No evidence of large-effect *trans*-acting mediators of the pervasive environmental influence was found; instead, genetic and environmental factors acted in a largely additive manner.

The human transition from pastoral and rural to urban lifestyles has been accompanied by an increase in the incidence of numerous chronic diseases such as asthma, diabetes and cancer[1]. Environmental contributors, which are likely to include dietary shifts, pollution and psychological factors, are the subject of continuing epidemiological research. It is equally interesting to determine whether genetic influences on disease susceptibility change across environments.

Because disease risk is commonly thought to involve differential gene expression[2], we have assessed the robustness of transcript abundance to environmental variation by performing a genome-wide association study (GWAS) on leukocyte gene expression profiles across two ancestries in three locations. Previously, we demonstrated that environmental geography[3] has a substantial effect on gene expression in Moroccan Amazigh individuals; here, we add the contrast with people of Arab descent, enabling us to test whether geography and/or ancestry affects each of several hundred robust associations between genotype and transcript abundance.

RESULTS

Population structure of southern Morocco

The Souss region in southern Morocco is home to several million people of two dominant ancestries who live in either cities or rural villages (**Fig. 1**). The Amazigh Berbers are descendants of the first modern humans who populated north Africa 35,000 years or more ago[4], and many still live in traditional villages in the low Atlas Mountains. The Arabs, by contrast, moved into southern Morocco between the seventh and eleventh centuries and tend to occupy lowland villages. The cities are inhabited by both groups, often retaining their linguistic and cultural identities.

In June and July of 2008, we collected peripheral blood samples from 284 healthy adults from four locations, including approximately equal numbers of men and women, and of Amazigh and Arabs. Half of the sample was from two high-density, low- to middle-income, urban communities, Anza and Dchiera, located on either side of the city of Agadir. The other half was from two rural villages near Tiznit, 120 km to the south. Boutroch is predominantly Amazigh and remains relatively isolated, whereas Ighrem is predominantly Arab and (on the basis of self-reported information and our observations at the collection site) many of the men, in particular, commute into the cities.

Leukocytes were isolated from serum, platelets and erythrocytes at the time of blood sampling by depletion filter technology[5] and fixed in RNALater solution within minutes of blood collection. Gene expression profiles were obtained from 208 high-quality RNA samples by using Illumina HumanHT12 bead arrays that include 48,804 probes, of which 22,300 RefSeq probes for 16,738 genes were deemed to have signal above background. To minimize batch effects, all samples were processed in the same week, and the extraction, labeling and hybridization steps were performed in accordance with a randomized block design. Whole-genome genotypes were obtained from whole-blood samples by using Illumina Human 610-Quad arrays. After quality control filters were applied, 516,972 SNPs were available for 194 of the individuals who also had gene expression profiles.

Population structure was assessed by examining the principal components of the variance of the genotype profiles using Eigenstrat

[1]Department of Genetics, North Carolina State University, Raleigh, North Carolina, USA. [2]SAS Institute Inc., Cary, North Carolina, USA. [3]Institute for Genome Science and Policy, Duke University, Durham, North Carolina, USA. [4]Queensland Institute of Medical Research, Brisbane, Queensland, Australia. [5]School of Biological Sciences, University of Queensland, Queensland, Australia. [6]HRH Prince Sultan International Foundation for Conservation and Development of Wildlife, Agadir, Morocco. Correspondence should be addressed to G.G. (ggibson.uq@gmail.com).

Received 6 July; accepted 13 October; published online 6 December 2009; doi:10.1038/ng.495

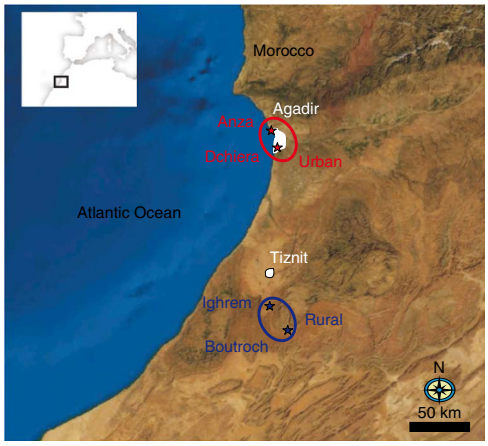

Figure 1 Map of the Souss region of southern Morocco showing the sampling locations. The two rural villages, Boutroch and Ighrem, are near the town of Tiznit. The urban locations, Anza and Dchiera, are north and south of the city of Agadir, respectively.

software[6]. Initial examination revealed several clusters of siblings and other close relatives (cousins or similar), whose similarity skewed the axes; where data were available, these identities were in agreement with participant records. After removal of these relatives, analysis of 163 unrelated individuals revealed seven significant eigenvectors (or genotypic principal component axes, gPCs). None of these explained more than 5% of the variance, and gPC3–gPC7 were heavily weighted by large clusters of SNPs on one or a few chromosomes. Such axes are commonly observed and do not provide reliable genome-wide estimates of population structure[7,8], but notably gPC3 distinguishes Ighrem from the other locations (**Supplementary Fig. 1a**).

A plot of the first two eigenvectors highlights the main historical influences on population structure in southern Morocco (**Fig. 2a**). gPC1 separates only a dozen individuals, and we inferred that this axis represents a sub-Saharan African contribution, consistent with expected levels of admixture in Morocco, by performing an analysis including 21 Yoruban individuals (**Supplementary Fig. 1b**). gPC2 is highly correlated with both location and self-reported ancestry; thus, we inferred that it captures the main component of Arab-Amazigh ancestry.

An unexpected aspect of this analysis is the positioning of Ighrem Arabs between Boutroch Amazigh and half of the Agadir Arabs along gPC2. Structure analysis[9] of 16,000 randomly chosen autosomal SNPs assuming admixture of two ancestral populations (**Fig. 2b**) confirmed that Ighrem residents tend to be a mixture, whereas most Amazigh are derived from one population, and only a few Agadir Arabs represent the other. Thus, there has probably been considerable admixture between these two groups over an extended period of time, possibly with recent movement of Arabs from other locations into Agadir. A slight shift of Ighrem Arabs toward the Amazigh pole of gPC2, relative to Agadir Arabs, would also be consistent with genetic exchange between the villages over 50 generations. Further sampling of villages in the region may reveal subtle population structure across southern Morocco[10–13].

Regional differentiation in gene expression

Next, we tested whether region, location and ancestry affect gene expression profiles, and if they do so in a gender-specific manner. Because location and ancestry are confounded in the villages, several parallel analyses were undertaken to tease apart these influences. Transcript abundance data were transformed by median centering on the $\log_2$ scale (**Supplementary Fig. 2**), which results in maximal overlap of profiles without altering their variance.

Gene-specific analysis of variance (ANOVA)[14] with expression as a function of region, gender and their interaction identified 1,521 probes that were significant at a false discovery rate (FDR) of 1% ($P < 0.0007$). Region, namely the rural (Boutroch plus Ighrem) versus city (Anza plus Dchiera) comparison, is the main effect in this joint analysis. Almost 7% of all expressed genes differentiate these individuals by this conservative criterion, whereas considerably less than 1% of the probes show gender differences (see **Supplementary Table 1** for a full list of genes). Among several classes of genes overrepresented in this lifestyle comparison, small nucleolar RNA genes stand out: 5 of the top 8 overall and 15 of 29 members of the SNORD family are in the highly significant list, as compared with only 1 of 10 SNORA genes. There is little in the literature to indicate why this is the case or what the physiological consequences may be, but epigenetic modification has been observed for many small nucleolar RNA genes[15].

Even more differentiation was observed when we fitted ANOVA models including location, gender and their interaction. Because exploratory analyses indicated that the Anza and Dchiera samples are indistinguishable for either gene expression or genotype, these samples were combined into a single location, Agadir, in all subsequent analyses. In the three-way comparison, 8,459 probes (38%) were significant at the 1% FDR threshold for location (**Supplementary Table 2**). Boutroch differs from Ighrem and also Agadir at over 7,000

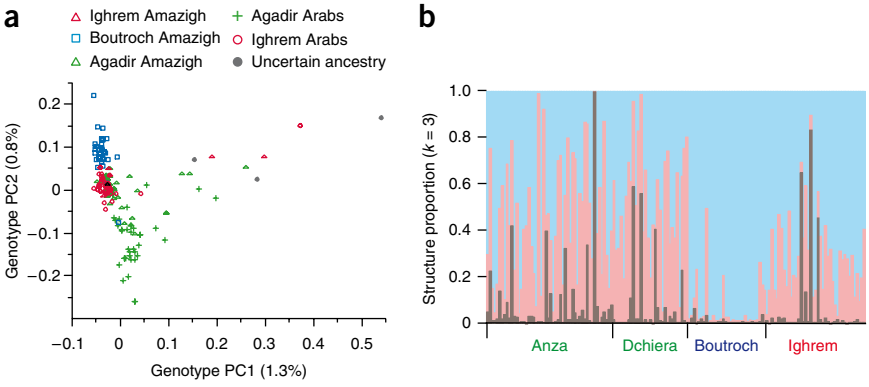

Figure 2 Population structure in southern Morocco. (**a**) Eigenstrat principal component analysis of 579,144 SNPs revealed seven significant eigenvectors, the first two of which, explaining only 1.3 and 0.8% of the genotypic variance, respectively, are plotted here. The ancestries were obtained by self-report. The three individuals with uncertain ancestry, possibly including sub-Saharan heritage, have high values of gPC1, which is characteristic of Yoruban ancestry (**Supplementary Fig. 1b**). (**b**) Structure analysis of 16,000 autosomal SNPs, assuming three populations ($k = 3$) and using an admixture model with correlated allele frequencies, highlights the same individuals with large gPC1 values (brown bars) and shows that Boutroch Amazigh are predominantly derived from one population group (light blue), whereas all other samples are a mixture of the two populations, represented by red and blue bars.

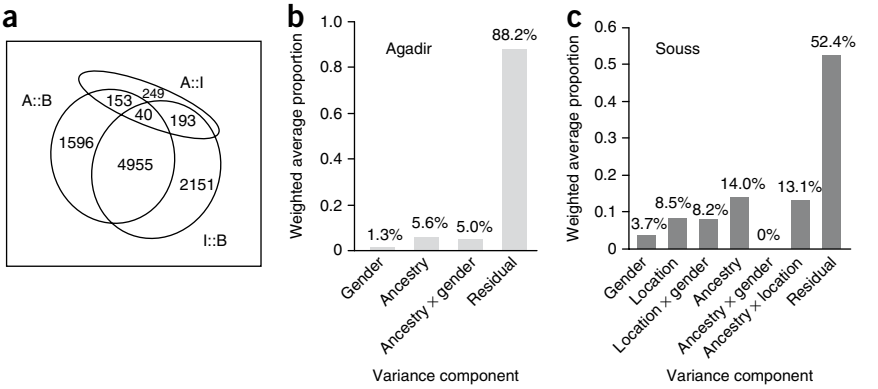

Figure 3 Location affects gene expression across the transcriptome. (**a**) Venn diagram of the number of genes significant at 1% FDR for ANOVA of the three pair-wise comparisons indicated. (**b**,**c**) Variance components of expression variation in the 118 residents of Agadir (**b**) (excluding nine individuals with strongly positive gPC1 scores, and including reassignment of ancestry according to gPC2 for only 11 individuals relative to self-report; **Supplementary Table 5**), where ancestry is modeled as gPC2 of the genotype variation as in **Figure 1a**, or for all 22,300 probes in the full sample of 208 individuals (**c**).

probes in each contrast, with a high degree of overlap (**Fig. 3a** and **Table 1**). Ighrem and Agadir are much more similar to one another, in part because there is considerably more diversity in the Ighrem sample that reduces the significance of the location contrast. Women are much more differentiated among locations than men (**Table 1**). These results confirm our previous report of substantial differentiation between Bedouin nomads, urban Anza and another remote Amazigh village, Sebt Nabor[3].

To evaluate the possible independent contribution of ancestry more carefully, we carried out variance component analysis of the expression variation. In Agadir alone, neither ancestry (modeled as the second eigenvector of the genotype data, gPC2) nor gender has a noteworthy impact on the principal components of the expression variation (**Fig. 3b**). In the total data set, however, there is evidence of a contribution: when fitted jointly with location, the ancestry and ancestry- and gender-by-location interaction terms make a substantial contribution to the expression profiles (**Fig. 3c**).

Although gender and ancestry affect the expression of fewer genes as compared with location, the plot of expression PC1 by PC2 for the most differentially expressed 1,500 genes indicates that for many genes the interaction between these three factors is complex (**Fig. 4**). This complexity is also seen in the expression profiles of characteristic individual genes (**Supplementary Fig. 3**). In general, Boutroch and Ighrem villagers separate along PC1, whereas high values of PC2 are obtained for all Boutroch residents (cluster 1) and for Arab women in Ighrem (cluster 2). Amazigh women from Ighrem (cluster 3) and the Ighrem men (cluster 4) have lower values of PC2, similar to those observed for all Agadir residents. The simplest interpretation is that cultural or behavioral differences, probably including time spent outside the village, contribute strongly to the observed gender and ancestry effects. Deeper sampling would be required to establish firmly whether intrinsic biological differences between the sexes and/or populations also make significant contributions to expression divergence in lymphocytes, as they appear to do for lymphoblast cell lines grown in culture[16–19].

Two classes of genes stand out as significantly differentially expressed among locations: namely, those encoding ribosomal proteins of the small and large subunits, as well as the cytoplasmic and mitochondrial compartments; and those encoding proteins involved in oxidative phosphorylation, which are highly upregulated in half of the Agadir residents (**Supplementary Fig. 4a**). All of the transcripts encoding these proteins form a module of co-regulated genes, but notably this module is not coexpressed with the SNORD family, which tends to be relatively downregulated in Agadir individuals but particularly highly expressed in the Arab women from Ighrem (**Supplementary Fig. 4b**). These differences may reflect differential abundance of leukocyte cell types, but ribosomal biosynthesis is also related to response to viral infection, and seems to be involved in tumorigenesis in conjunction with mitochondrial activity[20,21]. Oxidative phosphorylation is correlated with renal health and the production or disposal of free radicals[22]; thus, our data suggest that deeper evaluation of health risks associated with lifestyle transitions may be revealing.

Genome-wide association with gene expression variation

The genetic contribution to expression variation was evaluated by genome-wide association with expression of all 22,300 probes. Starting with a simple test of the correlation between each transcript abundance and each genotype, and filtering to retain only eSNPs with a minor allele frequency of >0.05, we observed 3,430 associations at $P < 10^{-8}$. Further filtering of eSNPs to retain only autosomal associations with annotated genes, and imposing the additional stringency of $P < 10^{-11}$ for putative *trans* associations between an eSNP on one chromosome and a probe on another chromosome, reduced this number to 1,636 associations: 1,569 (96%) of these associations are intra-chromosomal linkages and most are within 50 kb and hence *cis*-acting (**Supplementary Fig. 5**); only three are clearly in different

Table 1 Number of transcripts significant at 1% FDR

Location	ANOVA[a]	ANCOVA	Gender	ANOVA	ANCOVA	Interaction	ANOVA	ANCOVA
3-way	8,459	7,057	Male:female	151	233	Location × gender	133	203
Aga: Bou	6,744	4,974	Agadir	24	24	Fem (Aga: Bou)	4,830	3,791
Aga: Igh	635	651	Boutroch	13	14	Fem (Aga: Igh)	1,451	1,467
Bou: Igh	7,339	6,286	Ighrem	589	890	Mal (Aga:Bou)[b]	407	806
Aga: rural	1,521	607				Mal (Aga: Igh)	8	8

[a]ANOVA includes terms for location, gender and location × gender interaction. The FDR was evaluated by the conservative Benjamini and Hochberg method[40]. ANCOVA is the same model with an additional continuous covariate for ancestry (gPC2). Shown are the number of genes significant at the 1% FDR threshold for location effects (in a three-way comparison of Agadir (Aga), Boutroch (Bou) and Ighrem (Igh); between pairs of locations, or between Agadir (Aga) and the two rural sites combined); for gender (male versus female) effects (in the total sample or in each location individually); and for interaction effects (in the total sample, or between Agadir and either village for females or males separately). [b]The significance of this contrast was reduced by the small sample of Boutroch males (12 versus 26 females).

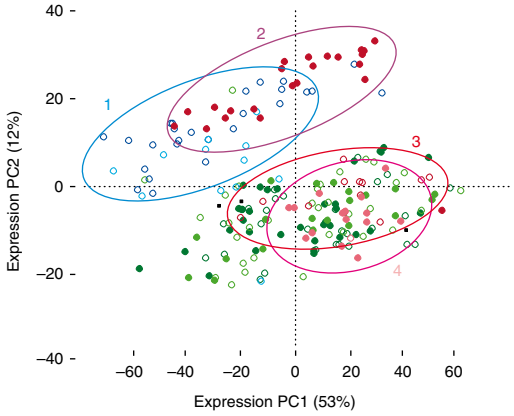

Figure 4 Principal component plot for the most differentially expressed genes. The two main principal components of the expression of the 1,500 most significant genes show significant separation of individuals by location (PC1 and PC2) and gender (PC2; all $P < 0.0001$). Individuals from Boutroch are blue, Ighrem are red and Agadir are green. Arabs are indicated with filled circles, Amazigh are open circles and males are lighter symbols for each color. Boutroch and Arab women from Ighrem (clusters 1 and 2) separate from Amazigh women and Arab men from Ighrem (clusters 3 and 4), who are closer to Agadir residents. If Boutroch residents and Ighrem Arab women are grouped and contrasted with Agadir residents, Ighrem Amazigh women and Ighrem men, 8,239 genes are significantly differentially expressed at the 1% FDR rate, more than any pair-wise comparison of locations. A similar plot for all genes is shown in **Supplementary Figure 11**.

chromosomal intervals. Facsimile associations were observed for 39 of the target genes represented by a second probe (37 *cis*, 2 *trans*). Reducing the data set further to exclude linked associations within haplotype blocks left 346 unique *cis* and ten unique *trans* associations at the stringent genome-wide significance level of 5%. These proportions are in good agreement with most other GWAS expression studies on blood or lymphocyte cell lines[16,17,23–26], and a 30-fold or greater excess of *cis* over *trans* associations is also supported by 1% FDR estimates of 600 and 20 genes, respectively (see **Supplementary Table 3** for a complete list of peak *cis* and *trans* associations).

Given the high degree of population structure for gene expression, we addressed the possibility that differentiation of eSNP allele frequencies may contribute to the associations observed by estimating the fraction of variation within subpopulations (F_{ST}) for each pair-wise comparison of location for the 516,972 SNPs and 16,500 of the genes. No fixed differences were observed, and plots of the F_{ST} comparisons (**Supplementary Fig. 6a**) indicate only moderate overall genetic differentiation with a few SNPs having F_{ST} values between 0.12 and 0.3. There is no tendency for these outliers to have increased differentiation in expression, and in fact almost all of the top 10% most differentially expressed genes are among the least genetically differentiated. Nor is there any correlation between F_{ST} and significance of gene expression divergence (**Supplementary Fig. 6b**), confirming that the expression differences observed between locations are for the most part not attributable to gene-specific allelic frequency differences between locations.

The robustness of the 3,430 associations to environmental sources of variance and population structure was further evaluated by fitting two additional linear trend models to the data. The first included location, gender and the interaction between them. The second included two measures of ancestry (the first three genotype eigenvectors and a four-way categorical ancestry cluster, see Online Methods), a matrix of relatedness based on an identity-by-descent measure[27], and gender interactions with ancestry cluster and genotype. **Figure 5a**,**b** shows the Manhattan plot of associations by chromosomal location for the second of these models, and the *cis-trans* plot of target against eSNP location, respectively. The logarithm of the genotype significance term is highly correlated ($r > 0.95$) between both of these models and the original correlation test (**Fig. 5c** and **Supplementary Fig. 7**). In addition, there is no evidence for significant genotype-by-location interactions in any of the association trend tests (**Fig. 5d**). Neither the ancestry nor the relatedness variance components explain an appreciable amount of the expression variation for any of the transcripts (**Supplementary Fig. 8**).

The absence of interaction effects can be visualized by plotting expression as a function of genotype, with color coding of each location, for each association. An example of a *trans* association in **Supplementary Figure 9** shows the clear trend of increased expression of *AMY1A* (chromosome 1) in homozygotes for the A allele of *ACTG1* gamma actin (chromosome 17) consistently across the three locations despite slight overall location effects. Expression of *AMY1A* is highly

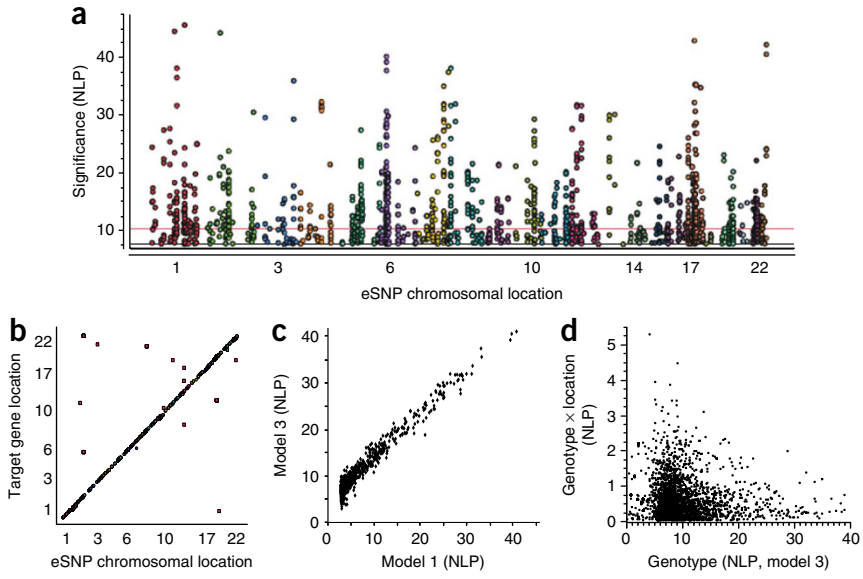

Figure 5 Genome-wide association with transcript abundance. (**a**) Manhattan plot of all 1,636 genome-wide associations at $P < 10^{-8}$ (negative log of the *P*-value, NLP > 8) for model 3, which includes control for genotype-determined ancestry, location, relatedness and gender. Each chromosome is indicated by a different color. The horizontal red line indicates the genome-wide significance threshold (NLP > 11.4) for *trans* associations. Note the excess of peaks at the major histocompatibility complex (MHC) complex on chromosome 6 owing to multiple *cis* eSNPs. (**b**) *Cis-trans* plot, showing target transcript location against eSNP location, indicating that most eSNPs are in *cis* to the regulated transcript, whereas only 13 *trans* associations at NLP > 11.4 are visible. (**c**) High correlation of significance measures for all eSNPs detected by simple correlation of genotype with expression (model 1) or robust control for ancestry, gender and location (model 3). (**d**) Absence of genome-wide significance for the genotype-by-location interaction effect, which is not correlated with the genotype effect.

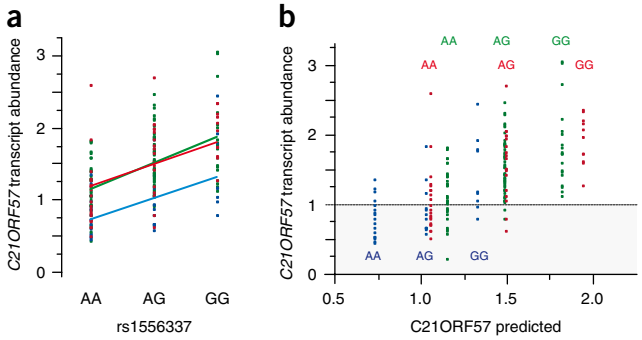

Figure 6 Relationship among genotype, expression and phenotype. (**a**) Typical example of a transcript (*C21ORF57*, encoding a putative metalloproteinase) that shows both a significant difference between locations ($P < 10^{-5}$) and a *cis* eSNP association with rs1556337 ($P < 10^{-13}$), but no interaction effect in an additive model on the $\log_2$ scale. Expression is lower in Boutroch (blue), whereas genotype has a consistent effect across all three locations (Ighrem, red; Agadir, green). (**b**) The actual versus predicted plot separates the genotypes by location for clarity. Suppose that a disease or phenotype is seen only in individuals with a transcript abundance of <1.0 (on a relative $\log_2$ scale), indicated by the gray area. Then, in Agadir and Ighrem almost all affected individuals are AA homozygotes, whereas in Boutroch heterozygotes and some GG homozygotes are also affected. There is thus a gene-by-environment interaction for the phenotype in the absence of a gene-by-environment interaction for transcription because the environment shifts more individuals into the susceptible zone. Similar arguments would apply to phenotypes with high expression values and to graded rather than threshold-dependent traits.

correlated with that of *AMY1B* ($r > 0.8$) and many other genes in a coexpression module, but the eSNP regulates only *AMY1A*, because it increases expression of the gene twofold in an additive manner. A similar plot for another representative gene (*C21ORF57*) shows highly significant location and genotype effects in *cis* (**Fig. 6a**) (see **Supplementary Fig. 9c** for further examples).

Novel associations with potential disease alleles

Expression associations detected in one tissue can identify regulatory variants that may be active in other tissues that are directly engaged in the etiology of disease[23,25,26]. For example *cis* linkages in peripheral blood are associated with the type 1 diabetes (T1D) susceptibility locus at chromosome 12q13. The strongest expression association is with transcription of the ribosomal protein gene *RPS26*, and network analyses have been used to argue that this gene is a more likely candidate for diabetes than is the initially reported[28] gene *ERBB3*. However, the strongest T1D association involves a SNP that differs from that associated with expression and/or splicing[24] of *RPS26*. We further found that the same linkage group of eSNPs, centered on rs10876864 in the *SUOX* gene 35 kb from *RPS26*, is also associated in *trans* with other RP26 paralogs (probably owing to cross-hybridization), and with *CCDC4* on chromosome 4, albeit at the suggestive significance level of $P = 3.5 \times 10^{-10}$. Intriguingly, expression of *RPS26* is only weakly correlated with that of the module of ribosomal proteins that differentiate locations (**Supplementary Fig. 4b**); therefore, this association does not contribute to the environmental effect on transcription of ribosomal protein genes.

Another *trans* association involves rs11987927 in *MYOM2* at 8p23, which interacts with the zinc finger transcription factor gene *ZNF71* at 19q13 and also with its own *MYOM2* transcript. Logic would suggest that the *cis* association probably affects the abundance of the MYOM2 myomesin protein, which in turn regulates *ZNF71*; however, the *trans* association is significantly stronger, and conditional dependence analysis[29,30] points in the opposite direction — that is, the *MYOM2* regulatory site influences *ZNF71*, which then feeds back on the *MYOM2* transcript (**Supplementary Fig. 10**). This example may be a cautionary tale concerning the interpretation of conditional dependence results. Notably, four of the seven strongest *trans* associations involve regulation by loci that include genes encoding structural proteins; the others are the laminin gene *LAMA5* (20q13) with the oxysterol binding protein gene *OSBPL2*, and the plekstrin homology domain gene *PLEKHM1* (17q21) with the kinase gene *MAPK8IP1*.

One further *trans* association deserves attention. Prolongation of fetal gamma hemoglobin expression in adults is often observed in individuals with thalassemia. We found association of two probes that detect transcripts of the hemoglobin genes *HBG1* and *HBG2* at 11p15 with rs766432 in the second intron of the zinc-finger proto-oncogene *BCL11A* at 2p16. This same SNP has been associated with the fraction of erythrocytes that contain measurable fetal hemoglobin[31], and alteration of BCL11A activity has been shown to drive differences in globin switching between mice and humans[32]. Another SNP in *BCL11A*, rs4671393, has been associated with abundance of two *BCL11A* transcript isoforms in the CEU (CEPH Utah residents with ancestry from northern and western Europe sample) and YRI (Yoruba in Ibadan, Nigeria) HapMap lymphoblast cell lines[33], but is not associated with *BCL11A* transcript abundance in our leukocyte data, suggesting that regulation of BCL11A translation or protein activity is more likely to be affecting *HBG1* and *HBG2* expression in our sample.

Numerous *cis* associations are likely to be of interest. We scanned the GWAS association database for overlap between our study and established disease associations at $P < 10^{-5}$. Of 1,628 entries, ten involve *cis* associations observed in our data set that explain between 15 and 55% of the transcript variance (**Supplementary Table 4**). Five of the associations are with disease conditions (rheumatoid arthritis, celiac disease, T1D, ulcerative colitis and systemic lupus erythematosus) and five are with endophenotypes (levels of the proteins PAFAH1B2 and ICAM-1, triglycerides, low-density lipid cholesterol and hip bone mineral density). The two serum protein associations[34,35] are with the same SNPs that we detected and hence suggest that protein abundance is largely regulated at the transcriptional level.

DISCUSSION

Genetic and environmental contributions to transcript variation

Our geographical genomic survey of gene expression variation in southern Morocco has highlighted two parallel and for the most part non-overlapping insights. On the one hand, it is evident that as much as half of the transcriptome is influenced by the environment in a highly coordinated manner such that where a person lives explains up to a quarter of the variation for a substantial fraction of the transcripts. The environmental influences are probably a combination of biotic and abiotic factors, in addition to cultural and behavioral ones, whereas genetic differences between the two north African ancestries are relatively minor. On the other hand, the genome is littered with strong genetic associations, mainly in *cis*, that explain between 15 and 60% of the variance of 5% of the transcripts. Impressive as these associations are, particularly because they are apparent in a sample of just under 200 individuals, they have essentially no bearing on most of the transcriptional variation and are not informative of the genetic basis of the environmental response.

The robustness of the associations observed to the environmental effect raises the issue of whether genotype-by-environment interactions influence the peripheral blood transcriptome at all. Genome-wide significant interaction effects are generally unlikely to occur in the

absence of significant main genotype effects[36]. The only circumstances in which they will occur are when the genotype effect is in the opposite direction in two locations, and if the genetic effect in these locations is at least the same magnitude as the main effects detected in this GWAS — in other words, if the effect can explain >30% of the variance of a particular transcript. Although a few such interactions may exist, it would take a study comparing several thousand individuals from each location to reveal weaker genotype-by-environment interactions. If the genetic architecture of transcription is similar to that of visible phenotypes such as height and body mass[37,38], then even such a study will be underpowered to explain most transcriptional variance.

A related issue is whether or not genotype-by-environment interactions at the level of transcription are necessary to explain genotype-by-environment interactions for disease. It is possible the small interactions beneath the level of detection of GWAS are prevalent, or alternatively that disease arises primarily as a result of rare alleles of major effect, whose penetrance may be modulated in an environment-specific manner. However, transcriptional interactions are not required to explain the increased incidence of chronic disease. It is not difficult to imagine that individuals that fall into the chief categories of transcriptome profiles (such as those implicated in **Fig. 4** and **Supplementary Fig. 4**) have different distributions of disease susceptibility that alter the genotype-disease association matrix across the genome, thereby inducing environment-by-genotype interactions for disease. Transcription of genes that contribute to this expression component may also correlate directly with disease, effectively uncovering cryptic variation and resulting in environment-specific eSNP disease associations without any interaction effect at the level of transcription[39] (**Fig. 6**). A corollary of this is that gene expression profiling might be used to stratify individuals at higher risk for disease, thereby increasing the resolution of GWASs by focusing attention on the subset of individuals in whom genetic effects on disease are most pronounced.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. NCBI GEO: Gene expression data from this study have been deposited under series GSE17065.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank all of the study participants in Agadir, Ighrem and Boutroch, and numerous individuals who facilitated sample collection, in particular the Idaghdour family. D. Ge and A. Motsinger-Reif provided timely computational support, and we also thank S. Biswas and J. Akey for providing HapMap genotypes. Funding for the study was provided by the University of Queensland. Y.I. was supported by a Fulbright Fellowship and G.G. by an ARC Australian Professorial Fellowship.

AUTHOR CONTRIBUTIONS

Y.I. collected the samples with the assistance of S.J.J. and processed them with K.V.S. and D.B.G.; K.M., S.H.L., D.B.G., P.M.V. and R.D.W. provided statistical and conceptual support for analysis of the data by Y.I., W.C., H.C.M. and G.G.; and Y.I. and G.G. conceived the study and wrote the paper. All authors read and contributed to the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Abegunde, D.O., Mathers, C.D., Adam, T., Ortegon, M. & Strong, K. The burden and costs of chronic diseases in low-income and middle-income countries. *Lancet* **370**, 1929–1938 (2007).
2. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10**, 184–194 (2009).
3. Idaghdour, Y., Storey, J.D., Jadallah, S.J. & Gibson, G. A genome-wide gene expression signature of environmental geography in leukocytes of Moroccan Amazighs. *PLoS Genet.* **4**, e52 (2008).
4. Arredi, B. *et al.* A predominantly Neolithic origin for Y-chromosomal DNA variation in North Africa. *Am. J. Hum. Genet.* **75**, 338–345 (2004).
5. Feezor, R.J. *et al.* Whole blood and leukocyte RNA isolation for gene expression analyses. *Physiol. Genomics* **19**, 247–254 (2004).
6. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
7. Fellay, J. *et al.* A whole-genome association study of major determinants for host control of HIV-1. *Science* **317**, 944–947 (2007).
8. Biswas, S., Scheinfeldt, L.B. & Akey, J.M. Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. *Am. J. Hum. Genet.* **84**, 641–650 (2009).
9. Pritchard, J., Stephens, M. & Donnelly, P. P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
10. Kéfir, R., Stevanovitch, A., Bouzaid, E. & Béraud-Colomb, E. Diversité mitochondriale de la population de Taforalt (12.000 ans bp - Maroc): Une approche génétique à l'étude du peuplement de l'Afrique du nord. *Anthropologie* **43**, 1–11 (2005).
11. Coudray, C. *et al.* Population genetic data of 15 tetrameric short tandem repeats (STRs) in Berbers from Morocco. *Forensic Sci. Int.* **167**, 81–86 (2007).
12. Ennafaa, H. *et al.* Mitochondrial DNA haplogroup H structure in North Africa. *BMC Genet.* **10**, 8 (2009).
13. Bosch, E. *et al.* Population history of North Africa: evidence from classical genetic markers. *Hum. Biol.* **69**, 295–311 (1997).
14. Wolfinger, R.D. *et al.* Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* **8**, 625–637 (2001).
15. Royo, H. & Cavaillé, J. Non-coding RNAs in imprinted gene clusters. *Biol. Cell* **100**, 149–166 (2008).
16. Dixon, A.L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* **39**, 1202–1207 (2007).
17. Stranger, B.E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224 (2007).
18. Cheung, V.G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365–1369 (2005).
19. Storey, J.D. *et al.* Gene-expression variation within and among human populations. *Am. J. Hum. Genet.* **80**, 502–509 (2007).
20. Kao, C.F., Chen, S.Y. & Lee, Y.H. Activation of RNA polymerase I transcription by hepatitis C virus core protein. *J. Biomed. Sci.* **11**, 72–94 (2004).
21. Ruggero, D. & Pandolfi, P.P. Does the ribosome translate cancer? *Nat. Rev. Cancer* **3**, 179–192 (2003).
22. Shah, S.V., Baliga, R., Rajapurkar, M. & Fonseca, V.A. Oxidants in chronic kidney disease. *J. Am. Soc. Nephrol.* **18**, 16–28 (2007).
23. Göring, H.H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.* **39**, 1208–1216 (2007).
24. Heinzen, E.L. *et al.* Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol.* **6**, e1000001 (2008).
25. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).
26. Heap, G.A. *et al.* Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med. Genomics* **2**, 1 (2009).
27. Hayes, B.J., Visscher, P.M. & Goddard, M.E. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* **91**, 47–60 (2009).
28. Schadt, E.E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6**, e107 (2009).
29. Chen, L.S., Emmert-Streib, F. & Storey, J.D. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol.* **8**, R219 (2007).
30. Rockman, M.V. Reverse engineering the genotype-phenotype map with natural genetic variation. *Nature* **456**, 738–744 (2008).
31. Menzel, S. *et al.* A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat. Genet.* **39**, 1197–1199 (2007).
32. Sankaran, V.G. *et al.* Developmental and species-divergent globin switching are driven by BCL11A. *Nature* **460**, 1093–1097 (2009).
33. Sankaran, V.G. *et al.* Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science* **322**, 1839–1842 (2008).
34. Melzer, D. *et al.* A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.* **4**, e1000072 (2008).
35. Paré, G. *et al.* Novel association of ABO histo-blood group antigen with soluble ICAM-1: results of a genome-wide association study of 6,578 women. *PLoS Genet.* **4**, e1000118 (2008).
36. Culverhouse, R., Suarez, B., Lin, J. & Reich, T. A perspective on epistasis: limits of models displaying no main effect. *Am. J. Hum. Genet.* **70**, 461–471 (2002).
37. Visscher, P.M. Sizing up human height variation. *Nat. Genet.* **40**, 489–490 (2008).
38. Soranzo, N. *et al.* Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size. *PLoS Genet.* **5**, e1000445 (2009).
39. Gibson, G. Decanalization and the origin of complex disease. *Nat. Rev. Genet.* **10**, 134–140 (2009).
40. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B.* **57**, 289–300 (1995).

ONLINE METHODS

Study population. Sampling was designed so that four localities representing two main lifestyles and including both genders were sampled, and both Arab and Amazigh ancestries were represented in each locality. Sampling of the two ancestries relied originally on self-reported information. The urban group consisted of residents sampled from two low-income districts, Anza and Dchiera, located seven miles apart on the north and south sides of Agadir, respectively. All of these individuals live a typical urban lifestyle characterized by a relatively dense human population, frequent traffic and the presence of industrial activities. The rural group consisted of villagers sampled from two sites, Ighrem and Boutroch, located 26 miles apart and 80 miles south of Agadir. Both villages are characterized by a traditional lifestyle based on agriculture and herding, but the villagers in Boutroch are more isolated and have very limited exposure to urban activities relative to the villagers in Ighrem. Obtaining samples from males from either village was challenging, and most of the males make occasional, or in some cases frequent, trips to neighboring cities. Boutroch is known to be a predominantly Amazigh village and is in the low Atlas mountains (latitude, 29.346; longitude, −9.368; altitude, 1335 m), whereas Ighrem is located in the foothills of the low Atlas mountains (latitude, 29.459; longitude, −9.672; altitude, 720 m) and is historically Arab with a small fraction of Amazigh residents; self-report confirmed these ancestry differences.

All study participants were between the ages of 18 and 50 yr, and the mean age of the three locations was similar (31–34 yr). The effect of age on gene expression was minimal; only 30 probes were significant at 1% FDR by ANCOVA with location and gender as fixed effects.

Collection protocol. The study was approved by the ethical review committees of the Moroccan Ministry of Health, North Carolina State University and the University of Queensland. Under informed consent, 284 peripheral blood samples were collected in the field; 215 and 209 of these samples were profiled for gene expression and genotype, respectively, but several were later discarded for quality control purposes (see below). The subjects reported that they were in good health at the time of sampling. Peripheral blood samples (~8 ml) were collected over the course of 6 d during the months of June and July 2008. The same collection protocol was followed for all samples to minimize heterogeneity due to technical reasons. All samples were collected within 4 h between 8:00 and 12:00. The total leukocyte population was isolated from ~6 ml, and within minutes its total RNA was stabilized by using a Leukolock Total RNA Isolation System[5] (Ambion). This system incorporates depletion filter technology to isolate leukocytes and to eliminate plasma, platelets and red blood cells and uses RNALater® to stabilize the RNA in the cells captured in the filter. The remaining blood was stored in EDTA tubes for DNA extraction. The filters and blood samples were kept on ice and then frozen at −45 °C within hours of collection at all study sites.

RNA and DNA preparation. Total RNA extraction, and cDNA and cRNA synthesis were performed with an Illumina TotalPrep RNA Amplification kit (Ambion) in accordance with the manufacturer's instructions. Total RNA samples were checked for quality with an RNA 6000 Nano LabChip kit and 2100 Bioanalyzer (Agilent). We retained 215 samples with high RNA quality (RNA integrity number > 8) for expression profiling. We extracted 209 DNA samples with a QIAamp DNA kit (Qiagen) and quantified them by using an ND-1000 instrument (NanoDrop Technologies). All DNA samples had 260/280 and 260/230 ratios of optical density within the range 1.70–2.05.

Gene expression profiling. HumanHT-12 beadchips (Illumina) were used to generate expression profiles of >48,000 transcripts by using 500 ng of labeled cRNA for each of the 208 samples in accordance with the manufacturer's recommended protocols. The order in which the samples were processed was randomized to minimize chip effects. The beadchips were hybridized and scanned with an Illumina BeadArray reader by K.S.'s laboratory at the Duke University Institute for Genomics and Science Policy (IGSP). The raw intensities were extracted with the Gene Expression Module in BeadStudio software (Illumina). Expression intensities were $\log_2$-transformed and median-centered by subtracting the median value of each array from each intensity value. This procedure preserves the variance of each sample, and inspection of the residuals indicated that they were reasonably distributed for ANOVA; in addition, an outlier filtering procedure provided further quality control. The top 22,300 transcripts with expression above background levels averaged across all of the arrays were retained for further analyses as described[3]. All array data have been submitted to GEO according to MIAME compliance guidelines and are available under accession number GSE17065.

Genome-wide genotyping. We assayed 209 samples with Infinium Human 610-Quad beadchips (Illumina) by following standard procedures, also at the Duke University IGSP. The Human 610-Quad SNP Chip contains over 610,000 markers based on HapMap release 23. The beadchips were imaged by using a BeadArray Reader (Illumina), and genotype calls were extracted with the Genotyping Module in BeadStudio software. Six samples with low intensity or a low call rate as assessed by the Illumina cluster measure (<95%) were removed, and all SNPs that had a call frequency of <99% were deleted. SNPs with a cluster separation value of <0.3 were checked manually, and those that could not be fixed manually were removed. Next, to screen for departure from Hardy-Weinberg equilibrium, we checked the quality of the raw and normalized data of autosomal SNPs with heterozygosity excess values between −1.0 to −0.1 and between 0.1 to 1.0, and any SNP cluster that was not clean was removed. The process of quality control checks resulted in retention of 579,144 SNPs in 203 individuals for the population structure analysis; this value was reduced to 516,972 for the association studies after removing SNPs with a minor allele frequency of <0.05.

Population structure, ancestry inference and F_{ST}. Principal component analysis (PCA) and a Bayesian approach were implemented in Eigenstrat[6] and Structure[9], respectively, to explore genetic structure among the samples. Relatedness between all pairs of individuals was estimated indirectly from identity by state measures using PLINK[41], and 65 of the individuals appeared to be related by virtue of having pi-hat scores of >0.125. We observed 15 pairs or triplets of full siblings (0.451 < pi-hat < 0.595, a range similar to that described for full siblings[42]), six clusters of lesser relatives (0.125 < pi-hat < 0.3) and four mixed clusters of 4–5 relatives of both types. By these criteria, 138 individuals did not appear to be related to any other individuals in the sample, and were combined with one randomly chosen member from each of the 25 clusters to result in 163 unrelated individuals for the population structure analysis. PCA was used to infer the extent of global genotypic variation in this set, retaining the first seven eigenvectors according to the Tracey-Widom test statistic. Close inspection of axes 3–7 indicated that they were dominated by a few SNPs that mapped to the same region of the genome (data available from the authors on request). The sub-Saharan contribution to PC1 was established by including matching genotypes for 21 Yoruban HapMap individuals (provided by J. Akey and S. Biswas, University of Washington) in an expanded analysis. Structure[9] was used to infer population structure with a subset of 16,000 autosomal SNPs (randomly selected and approximately uniformly distributed on the 22 autosomes) at $k = 2$–5 using the admixture model with correlated allele frequencies and 20,000 iterations after a burn-in length of 20,000.

Subsequently, relatedness was recalculated more formally[27] for all individual pairs by using $\hat{A}_{ij}$ averaged over $l = 1$ to n loci:

$$\hat{A}_{ij} = [\Sigma\,(x_{il} - 2p)\,.\,((x_{jl} - 2p)/2pq)]/n$$

where $x_{il} = 0$, 1 or 2 according to whether individual i has genotype aa, Aa or AA at locus l, p (q) is the allele frequency of A (a), and $2p$ is the mean of x_l.

F_{ST} estimates between locations were calculated for each of the 516,972 SNPs included in the association study by using PROC ALLELE in SAS version 9.2 (SAS Institute). This implementation uses the method of moments approach in an ANOVA framework and expected mean squares to estimate F_{ST}. The method assumes 'random' (in contrast to 'fixed') populations and accounts for common evolutionary history. Gene-specific F_{ST} estimates were calculated by averaging F_{ST} measures of all SNPs in each gene and in flanking 5′ and 3′ UTR regions. Plots of F_{ST} by SNP and gene show typical upper values of 0.08, 0.10 and 0.12 for comparisons of Agadir with Ighrem, Boutroch with Ighrem, and Agadir with Boutroch, respectively (**Supplementary Fig. 6a**). A few SNPs exceed these values, the maximum being 0.3: no fixed differences between the locations were observed. To test for a possible influence of divergence in allele

doi:10.1038/ng.495

and genotype frequencies on gene expression divergence between locations, we examined the correlation between F_{ST} and fold change in expression, or significance of differential expression for each pair-wise comparison. There was no relationship between these measures (P values for all correlations > 0.047, percentage variance explained < 0.1%), nor was there an excess of outliers with high F_{ST} and high expression divergence (**Supplementary Fig. 6b**). Genetic differentiation thus does not significantly contribute to the location effects.

Principal variance component analyses, ANOVA and ANCOVA. Principal variance component analyses were performed on gene expression data by using JMP Genomics v3.2 (SAS Institute). Expression principal components (ePCs) were modeled as a function of various effects, assuming that each is a random term. A series of models was used to partition variance components into different combinations of the following factors and their pair-wise combinations: location (or lifestyle), gender and gPC2 (the second principal component of the genotypic variance, corresponding to the Arab-Amazigh axis of diversity). The magnitude and significance of differential expression of individual transcripts were evaluated by ANOVA and analysis of covariance (ANCOVA) through JMP Genomics using PROC MIXED as implemented in SAS and incorporating an outlier removal algorithm with a 5% false positive rate criterion. The following ANOVA models were used for differential expression analysis:

$$\text{expression} = \mu + \text{location} + \text{gender} + \text{location} \times \text{gender} + \varepsilon, \text{or}$$

$$\text{expression} = \mu + \text{lifestyle} + \text{gender} + \text{lifestyle} \times \text{gender} + \varepsilon$$

and gPC2 was added as a covariate for ANCOVA. Location (Agadir, Ighrem or Boutroch), lifestyle (urban or rural) and gender (male or lemale) were considered fixed effects. The error ε was assumed to be normally distributed with mean zero.

A marked feature of the PCA of the total data set is the presence of such a strong correlation structure in the data that ePC1 explains 21% and ePC1–ePC5 combined explain 50% of the transcriptional variance. In addition, almost half (47.6%) of the variation captured by ePC1–ePC5 can be decomposed into effects of the Arab-Amazigh axis of variation (gPC2), location, gender, and pair-wise interactions among these factors (**Fig. 3c**). This analysis is described in detail in ref. 43. It is substantially in agreement with the gene-specific ANOVA, which revealed similar magnitudes of contribution of the various effects. Taken together, the two modes of analysis imply that genetic and non-genetic effects both contribute significantly to transcriptional variation in our human data set. In addition, to evaluate possible environmental effects on alternative splicing, we fitted a mixed model for each gene targeted by more than one probe in the array and found evidence for 245 transcriptome-wide significant ($P < 1.2 \times 10^{-5}$) location-specific differences in transcript isoform abundance (**Supplementary Note**).

The absence of a relationship between transcript size (and GC content) and significance of differential expression (**Supplementary Fig. 12**) shows that there is no tendency for shorter transcripts to be differentially expressed between locations or lifestyles, indicating that enrichment for short transcripts such as the SNORD gene family is not due to degradation or technical artifacts.

Clustering and functional enrichment annotation. Clustering was generated with Ward's method in JMP Genomics v3.2. The gene ontology and pathway analyses were generated through the use of Panther[44] and KEGG[45]. Genes whose expression was significantly differentially regulated were included by using stringent cutoffs as described in the Results. Enrichment analysis was used to calculate the probability that the number of genes in each biological function, pathway and/or disease assigned to that data set was greater or less than expected by chance given the numbers of genes expressed in the samples. Corrections for multiple testing were achieved using Bonferroni or Benjamini-Hochberg methods depending on the analysis.

Genome-wide association tests. Tests for association of gene expression levels with each genotype were performed by both ANOVA (to test for genotype effects irrespective of allelic trends) and regression (to test for a linear trend, where heterozygotes are intermediate in phenotype owing to additive allelic effects) as implemented in PROC MIXED with SNP as a class variable or continuous variable, respectively, using SAS 9.2 and JMP Genomics v3.2. First, the whole allelic data set was coded as 0, 1 or 2, where each number represents the number of copies of the minor allele. Each of 516,792 SNPs was tested for association with each of the 22,300 expressed transcripts. This analysis gave rise to a genome-wide Bonferroni threshold of 4×10^{-12} for *trans* associations (NLP > 11.4, which is likely to be conservative given the linkage disequilibrium (LD) structure across the genome) and, assuming that 200 common SNPs are in 100 kb of each transcript probe, a threshold of $0.05/(22300 \times 200) = 1 \times 10^{-8}$ for *cis* associations (this value is also likely to be conservative because the median number of linked SNPs is <100). Note that a small fraction of putative *cis* eSNPs are more distant from the transcription start site than 50 kb on either side. We pragmatically distinguished *cis* from *trans* effects by plotting the eSNP and probe coordinates for each chromosome. Only three associations on the same chromosome were clearly off the diagonal; the remainder were within 1% of the chromosome arm length of the target probe and operationally likely to be *cis*-acting. The 1% FDR threshold was estimated by using the relationship FDR = $m \times$ alpha/(number of positives at alpha), where m is the total number of comparisons. Assuming 10^6 independent *cis* tests and 2×10^9 independent *trans* tests allowing for LD, approximate 1% FDR thresholds were found with 600 and 20 associations, respectively, at $P < 6 \times 10^{-6}$ and $P < 10^{-10}$. Although the complex dependency structure of the genotype and expression data caution against too literal interpretation of these numbers, similar relative numbers of the two types of association are obtained with different assumptions about non-independence of the tests.

Tests of association were carried out with three models. First, we used the following basic correlation model, where μ is the mean measure of transcript abundance and the error ε is assumed to be normally distributed with a mean of zero:

$$\text{expression} = \mu + \text{SNP} + \varepsilon \,(\text{model 1})$$

The 10,000 most significant associations from this model were brought forward for two further analyses. Model 2 assessed the effects of location (Agadir, Ighrem or Boutroch) and gender (male or female):

$$\text{expression} = \mu + \text{location} + \text{gender} + \text{SNP} + \text{SNP} \times \text{location} + \text{gender} \times \text{location} + \varepsilon \,(\text{model 2})$$

We also accounted for location, ancestry, relatedness and gender in a third model:

$$\text{expression} = \mu + \text{location} + \text{gender} + \text{relatedness} + \text{gPC1} + \text{gPC2} + \text{gPC3} + \text{gCluster} + \text{SNP} + \text{SNP} \times \text{location} + \text{gender} \times \text{gCluster} + \text{gender} \times \text{location} + \varepsilon \,(\text{model 3})$$

where gPC1-3 correspond to genotypic principal component eigenvectors of axis 1, 2 and 3 computed with Eigenstrat; and gCluster represents clustered ancestry, where the 194 samples were clustered into four groups corresponding largely to Agadir Arabs, Ighrem Arabs, Boutroch Amazighs and admixed individuals from Agadir and Ighrem, which accounts for location in an unbiased manner relative to ancestry. Relatedness was fitted as a random effect. Considerable overlap was observed between our set of GWAS-significant hits and highly significant eSNP associations reported in four other expression GWASs on peripheral blood or its derivatives, depending on the stringency adopted (**Supplementary Note**).

41. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
42. Visscher, P.M. *et al.* Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.* **2**, e41 (2006).
43. Idaghdour, Y. *Genetic and Environmental Components of Human Leukocyte Gene Expression Variation in Morocco.* PhD thesis, North Carolina State Univ. (2009).
44. Thomas, P.D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141 (2003).
45. Okuda, S. *et al.* KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res.* **36**, W423–W426 (2008).