ONLINE METHODS

Study population. Sampling was designed so that four localities representing two major lifestyles were sampled, including both genders, while both Arab and Amazigh ancestries were represented in each locality. Sampling of the two ancestries relied originally on self-reported information. The urban group consists of residents sampled from two low income districts (Anza and Dchiera) seven miles apart, on the north and south sides of Agadir respectively. All of these individuals live a typical urban lifestyle characterized by relatively dense human population, frequent traffic, and the presence of industrial activities. The rural group consists of villagers sampled from two sites (Ighrem and Boutroch) 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 both villages 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 meters), whereas Ighrem is located in the foothills of the low Atlas mountains (Latitude = 29.459, Longitude -9.672, Altitude: 720 meters) and is historically Arab with a small fraction of Amazigh residents; self-report confirmed these ethnic differences.

All study participants were between the ages of 18 and 50, and the mean age of the three locations was similar (31 to 34). The effect of age on gene expression was minimal with just 30 probes significant at the 1% FDR level from 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. A total of 284 peripheral blood samples were collected in the field under informed consent, 215 and 209 of which were gene expression and genotype profiled, respectively, but several samples were later discarded for quality control purposes as described 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 six days during the months of June and July 2008. The same collection protocol was followed for all samples in order to minimize heterogeneity due to technical reasons. All samples were collected within four hours between 8:00am and noon. The total leukocyte population was isolated from ~6 ml

and within a few minutes its total RNA was stabilized using the Leukolock® Total RNA Isolation System (Ambion, Austin TX)[5]. The system incorporates depletion filter technology to isolate leukocytes and eliminate plasma, platelets, and red blood cells, and RNA*Later*® 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 -45ºC within a few hours of collection at all study sites.

RNA and DNA preparation. Total RNA extraction, and cDNA and cRNA synthesis were performed with the Illumina TotalPrep RNA Amplification kit (Ambion, Austin TX) following the manufacturer's instructions. Total RNA samples were quality checked with the RNA 6000 Nano LabChip® kit and the 2100 Bioanalyzer (Agilent, Santa Carla CA). 215 samples with high RNA quality (Agilent's RNA Integrity Number RIN > 8) were retained for expression profiling. 209 DNA samples were extracted with QIAamp DNA Kit (Qiagen, Germantown MD) and quantified using the ND-1000 (NanoDrop Technologies, Wilmington DE). All DNA samples have 260/280 and 260/230 ratio of optical density within the range 1.70-2.05.

Gene expression profiling. Illumina's HumanHT-12 beadchips were used to generate expression profiles of more than 48,000 transcripts with 500ng of labeled cRNA for each of the 208 samples, again following 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 KS's laboratory at the Duke University Institute for Genomics and Science Policy (IGSP). The raw intensities were extracted with the Gene Expression Module in Illumina's BeadStudio software. Expression intensities were log2 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 indicates that they are reasonably distributed for ANOVA, noting that an outlier filtering procedure provides further quality control. The top 22,300 transcripts with expression above background levels averaged across all the arrays were retained for further analyses as previously described[3]. All array data has been submitted to GEO according to MIAME compliance guidelines and is available under accession number GSE17065.

Genome-wide genotyping. A total of 209 samples were assayed with Illumina's Infinium Human 610-Quad beadchip 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 using Illumina's BeadArray Reader and genotype calls were extracted with the Genotyping Module in Illumina's BeadStudio software. Six samples with low intensity or low call rate using the Illumina cluster measure (<95%) were deleted and all SNPs that had a call frequency below 99% were deleted. SNPs with a cluster separation value below 0.3 were checked manually and those that could not be fixed manually were deleted. Next, to screen for departure from Hardy-Weinberg equilibrium, the quality of the raw and normalized data of autosomal SNPs with heterozygosity excess values between -1.0 to -0.1 and 0.1 to 1.0 was checked and any SNP cluster that was not clean was deleted. The process of quality control checks resulted in retention of 579,144K SNPs in 203 individuals for the population structure analysis, and this was reduced to 516,972 for the association studies after removing SNPs with a minor allele frequency less than 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* [40] and 65 of the individuals appeared to be related by virtue of having *pi*-hat scores >0.125. 15 pairs or triplets of full siblings were observed (0.451< *pi*-hat <0.595, a similar range to that previously described for full sibs[41]), 6 clusters of lesser relatives (0.125<*pi*-hat<0.3), and 4 mixed clusters of 4-5 relatives of both types. 138 individuals do not appear to be related to any other individuals in the sample by these criteria, which combined with one randomly chosen member of each of the 25 clusters resulted 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 (TW) test statistic. Close inspection of axes 3 through 7 indicated that they are dominated by a small number of SNPs that map to the same region of the genome (data available on request). The sub-Saharan contribution to PC1 was established by including matching genotypes for 21 Yoruban HapMap individuals, provided by Josh Akey and Shameek Biswas, in an expanded analysis. *Structure* was used to infer population structure using 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 with 20,000 iterations after a burn-in length of 20,000.

Subsequently, relatedness was re-calculated more formally[27] for all individual pairs 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 allele frequency of A (a), and 2p is the mean of x_l.

F_{ST} estimates between locations were calculated by for each of the 516,972 SNPs included in the association study, using PROC ALLELE in *SAS 9.2* (SAS, Cary NC). This implementation uses the method of moments approach in an ANOVA framework using 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 within each gene and in flanking 5' and 3' UTR regions. Plots of F_{ST} by SNP and gene are provided in **Supplementary Figure 6a** online, which 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. Occasional 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 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. As shown in **Supplementary Figure 6b** online, there was no relationship between these measures (p-values for all correlations greater than 0.047, percent variance explained < 0.1%), and nor did we observe an excess of outliers with high F_{ST} and high expression divergence. 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 using *JMP Genomics* v3.2 (SAS Institute, Cary NC). Expression principal components (ePC) were modeled as a function of various effects, assuming that each is a random term. A series of models were 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 was evaluated using analysis of variance (ANOVA) and analysis of covariance (ANCOVA) through *JMP Genomics* (SAS Institute, Cary NC) using PROC MIXED as implemented in *SAS*, incorporating an outlier removal algorithm with a 5% false positive rate criterion. The following ANOVA models were used for differential expression analysis:

$$\text{Expression} = \mu + \text{Location} + \text{Gender} + \text{Location*Gender} + \varepsilon, \text{ or}$$

$$\text{Expression} = \mu + \text{Lifestyle} + \text{Gender} + \text{Lifestyle*Gender} + \varepsilon$$

and gPC2 was added as a covariate for the ANCOVAs. Location (Agadir, Ighrem, Boutroch), Lifestyle (Urban or Rural) and Gender (Male or Female) were considered fixed effects. The error ε was assumed to be normally distributed with mean zero.

A striking feature of the principal component analysis of the total dataset is the presence of such a strong correlation structure in the data that expression PC axis 1 (ePC1) explains 21% and ePC1 through ePC5 combined explain 50% of the transcriptional variance. Furthermore, almost half (47.6%) of the variation captured by ePC1 through ePC5 can be decomposed into effects of the Arab-Amazigh axis of variation (gPC2), Location, Gender, and pairwise interactions among these factors (**Fig. 3c**). This analysis is described in detail in Idaghdour, 2009[42]. 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 dataset. Furthermore, in order 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** online).

Absence of relationship between transcript size (and GC content) and significance of differential expression (**Supplementary Fig. 12** online) 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 (SAS Institute, Cary NC). The gene ontology and pathway analyses were generated through the use of *Panther* [43] and *KEGG* [44]. Genes whose expression was significantly differentially regulated were included using stringent cutoffs as described in the Results section. 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 is 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 using both ANOVA (to test for genotype effects irrespective of allelic trends) and regression (testing for a linear trend, where heterozygotes are intermediate in phenotype due to additive allelic effects) as implemented in PROC MIXED with SNP as a class or continuous variable respectively, using *SAS 9.2* and *JMP Genomics 3.2* (SAS). First, the entire 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 gives 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 LD structure across the genome), and assuming that 200 common SNPs are within 100kb of each transcript probe, to $0.05/(22300 \times 200) = 1 \times 10^{-8}$ for *cis*-associations (this is also likely to be conservative as the median number of linked SNPs is less than 100). Note that a small fraction of putative *cis*-eSNPs are more distant from the transcription start site than 50kb on either side. We pragmatically distinguished *cis*- from *trans*- effects by plotting the eSNP and probe coordinates for each chromosome. Just 3 associations on the same chromosome were clearly off the diagonal; the remainder are within 1% of the chromosome arm length of the target probe and operationally likely to be *cis*-acting. The 1% false discovery rate threshold was estimated using the relationship FDR = m*alpha / [# 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 $p<10^{-10}$. While the complex dependency structure of the genotype and expression data caution against too literal interpretation of these numbers,

similar relative numbers of the two type of association are obtained with different assumptions about non-independence of the tests.

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

$$Expression = \mu + SNP + \varepsilon \qquad \text{(Model 1)}$$

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):

$$Expression = \mu + Location + Gender + SNP + SNP^{*} Location + Gender^{*} Location + \varepsilon \qquad \text{(Model 2)}$$

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

$$Expression = \mu + Location + Gender + Relatedness + gPC1 + gPC2 + gPC3 + gCluster + SNP + SNP^{*} Location + Gender^{*}gCluster + Gender^{*}Location + \varepsilon \qquad \text{(Model 3)}$$

where gPC1-3 correspond to genotypic principal component eigenvectors of axis 1, 2 and 3 computed with *Eigenstrat*; and gCluster represents clustered ethnicity 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 ethnicity. 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-GWAS studies of peripheral blood or its derivatives, depending on the stringency adopted (**Supplementary Note** online).