Supplementary Information

Geographical Genomics of Human Leukocyte Gene Expression Variation in Southern Morocco

Youssef Idaghdour[1], Wendy Czika[2], Kevin V. Shianna[3], S. Hong 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] and Greg Gibson[5*]

1. Department of Genetics, North Carolina State University, Raleigh NC, USA
2. SAS Institute Inc., Cary NC, USA
3. Institute for Genome Science and Policy, Duke University, Durham NC, 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

Content

Supplementary Note
Supplementary Figures 1-12

Note: **Supplementary Tables 1-5** are provided as separate spreadsheet files on Nature Genetics website.

Supplementary Note

Concordance of probe effects and replication of associations. The 22,300 probes included in the analysis represent 16,738 annotated genes, since 5562 probes are duplicates that are designed to detect all (A), multiple (S), or isoform-specific (I) transcripts. Agreement between pairs of probes for the same gene with respect to the significance of the location effect at $P < 0.001$ was greater than 80%. Although the Illumina bead arrays are not specifically designed for detection of alternative transcripts, we further assessed whether there might be evidence for location-dependent alternative splicing. 973 of 4087 genes (24%) with multiple probes showed a significant Probe-by-Location effect ($P < 1.2 \times 10^{-5}$); some of this can be attributed to poor expression of one or more probes, but 245/1210 genes (20%) with all probes expressed at least two-fold above the grand mean (that is, with median centered expression values greater than 1) remain transcriptome-wide significant. We examined 10 highly significant cases in detail and found that removal of probes that only detect a subset of transcripts reduced the interaction effect substantially in 7 cases (*AKAP13*, *SLA*, *TAPBP*, *GSDML*, *ELF2*, *TCEAL8* and *MAPK14*), but for the other cases, all probes supposedly detect the same transcript (*UBQLN1*, *SEC24B* and *C20ORF121*). Further data mining revealed cases where disagreement between probes of the same type was observed but it should be noted that the A, S, I designations are not always reliable, due to ongoing computational and empirical re-annotation of the genome. Nevertheless, statistically speaking, our data provides evidence for several hundred instances of location-specific regulation of transcript isoform abundance. This may be attributed either to regulation of alternate splicing or of miRNA mediated alternate mRNA isoform degradation across locations, but alternate methods will be required to validate this conclusion more generally.

The uncertainty regarding whether discrepancies in probe reporting for the same gene had a technical or biological basis in each case precluded estimation of eSNP for alternate splicing in a systematic manner. However, we did observe that 10% of the robust eSNPs, including *cis* and *trans* associations, showed facsimile associations with a second probe from the same gene. This is greatly in excess of expectations if the associations were just due to noise since just 2% of transcripts show any evidence for association. Considerable overlap was also observed with highly significant eSNP associations reported in four other expression-GWAS studies of peripheral blood or its derivatives, depending on the stringency adopted. In Heinzen *et al*'s study[1] of 80 PBMC samples, 10/18 (56%) of the high confidence cis-

associations (NLP > 12) are observed in our set of GWAS-significant hits (and 43/361 exon-level associations). In Heap *et al*'s study[2] of peripheral blood from 115 Celiac patients, 68/179 (40%) of the genes at the same significance cutoff as we use (NLP > 8) are replicated. In Stranger *et al*'s study[3] of CEU HapMap lymphoblast cell lines, 39/174 (20%) of the top associations are replicated. In Göring *et al*'s linkage study[4] of Buffy coats (leukocytes plus platelets) from 1,240 individuals, 6/16 (37%) of the strongest cis QTL peaks are coincident with our significant cis-eSNPs (plus one *trans* association with MAPK8IP1), as well as 81/738 (11%) of the lower confidence (LOD>3) *cis*-linkages. In each of these studies, approximately one quarter of the associations are exact, namely involving the same SNP-transcript pair, and the remainder report the strongest association with a different SNP in the gene/linkage block. Given that these studies are performed with different types of blood sample, different genotyping and expression profiling platforms, and different analytical methods, the overlap is considerable, and in fact more extensive than previously reported comparisons, but it is not possible to estimate what proportion of the failures to replicate are due to biological differences, technical differences, or statistical power at genome-wide significance levels.

References

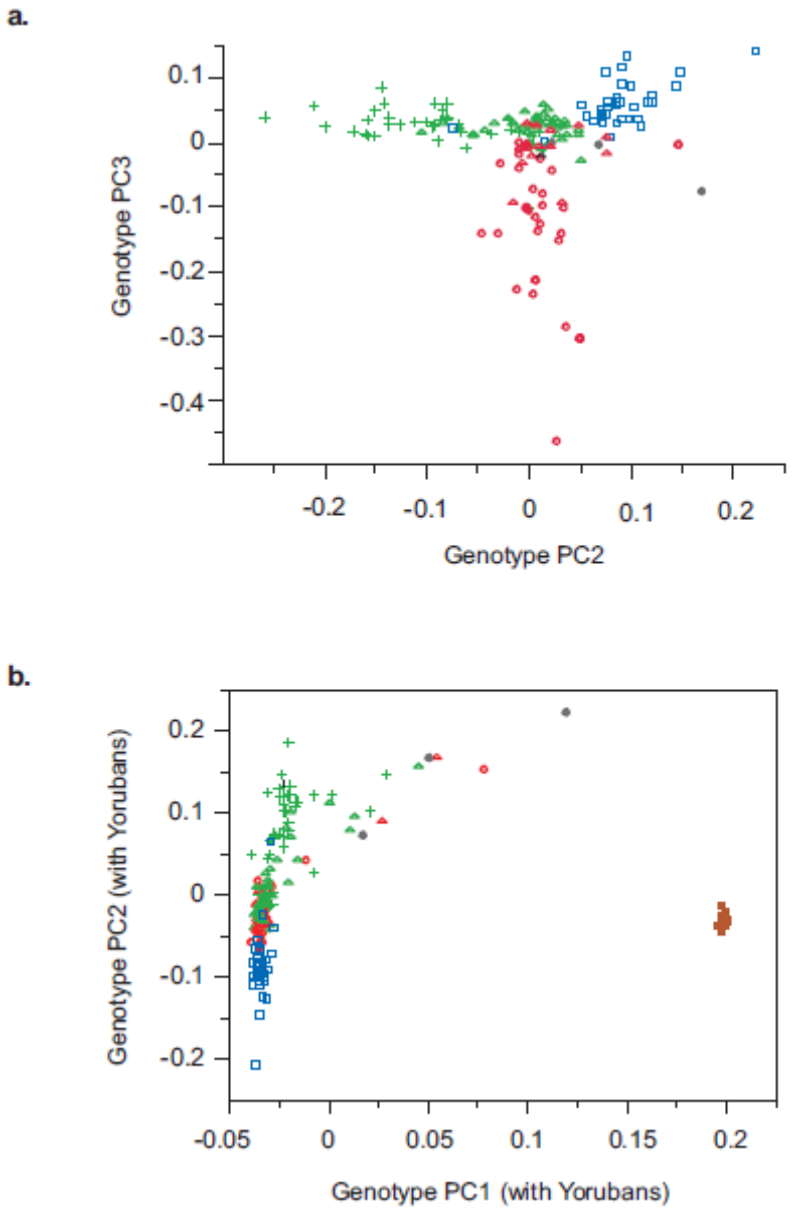
1. Heinzen, E.L. *et al*. Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol*. **6**, e1000001 (2008).

2. Heap, G.A. *et al*. Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med Genomics*. **2**, 1 (2009).

3. Stranger, B.E. *et al*. Genome-wide associations of gene expression variation in humans. *PLoS Genet*. **1**, e78 (2005).

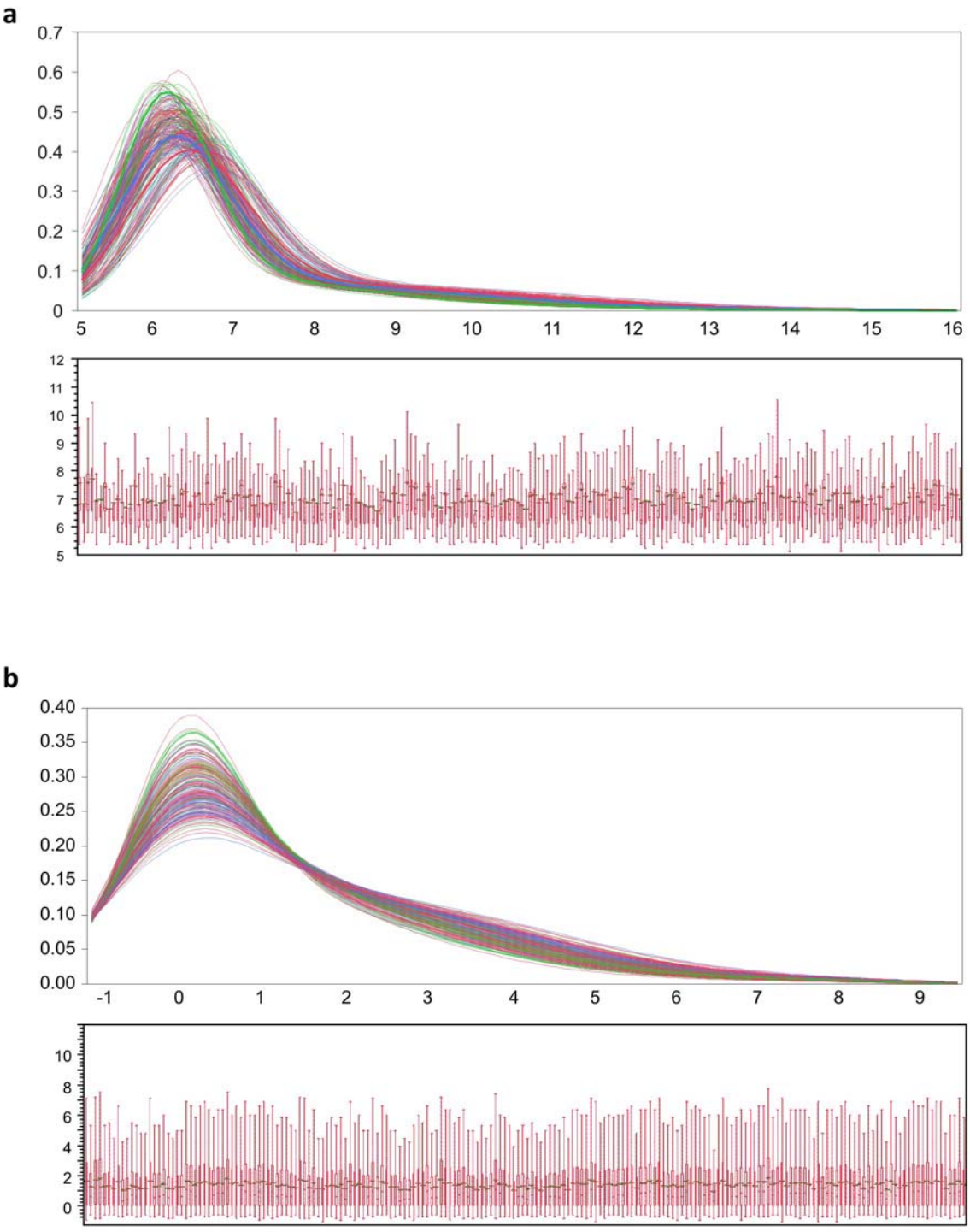
4. Göring, H.H. *et al*. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet*. **39**, 1208-1216 (2007).

Supplementary Figure 1. Expanded genotypic principal component analysis



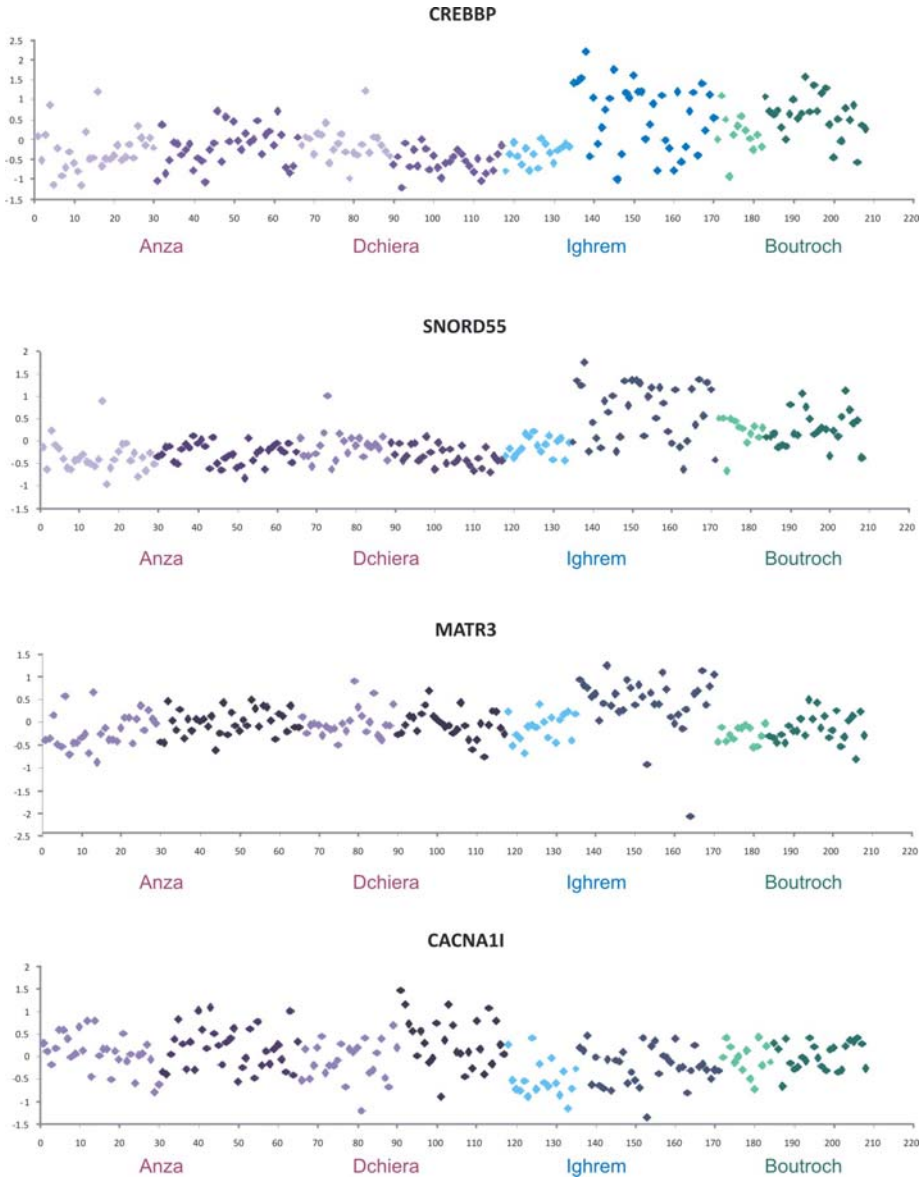
Supplementary Figure 1. (**a**) Plot of genotypic Principal Components 2 and 3 from Eigenstrat analysis of unrelated individuals (labeled by self-report), showing separation of Ighrem individuals (Arabs, red circles; Amazigh, red triangles) by gPC3. Study of the SNP loadings indicate that this signal is largely due to a small number of chromosomal regions. Also note the shift of Ighrem residents toward the Amazigh pole of gPC2 relative to Agadir Arabs (green plus signs) and Boutroch Amazigh (blue squares). Green triangles are Agadir Amazigh. (**b**) Inclusion of 21 Yoruban samples (brown filled squares) in an expanded analysis demonstrates that gPC1 represents a sub-Saharan axis of variation, since these individuals have extreme values of gPC1 along the same eigenvector as the dozen divergent individuals in **Figure 2a**. Three individuals who indicated that they were of uncertain ethnicity, thought to include sub-Saharan ancestry, are indicated as gray circles. 11 individuals whose ethnicity was reassigned from self-report to genotypic ethnicity for the analysis of within-Agadir variation in **Figure 3b** are indicated in **Supplementary Table 5**.

Supplementary Figure 2. Raw distributions of log base 2 Illumina scores for each sample



Supplementary Figure 2. Raw distributions of log base 2 Illumina scores for each sample (**a**) are transformed by median centering (**b**) to optimally reduce heterogeneity due to array effects without variance normalization. In this plot, Agadir residents are red, Ighrem blue, and Boutroch green. There is no significant tendency for differences in frequency distributions across the three populations.

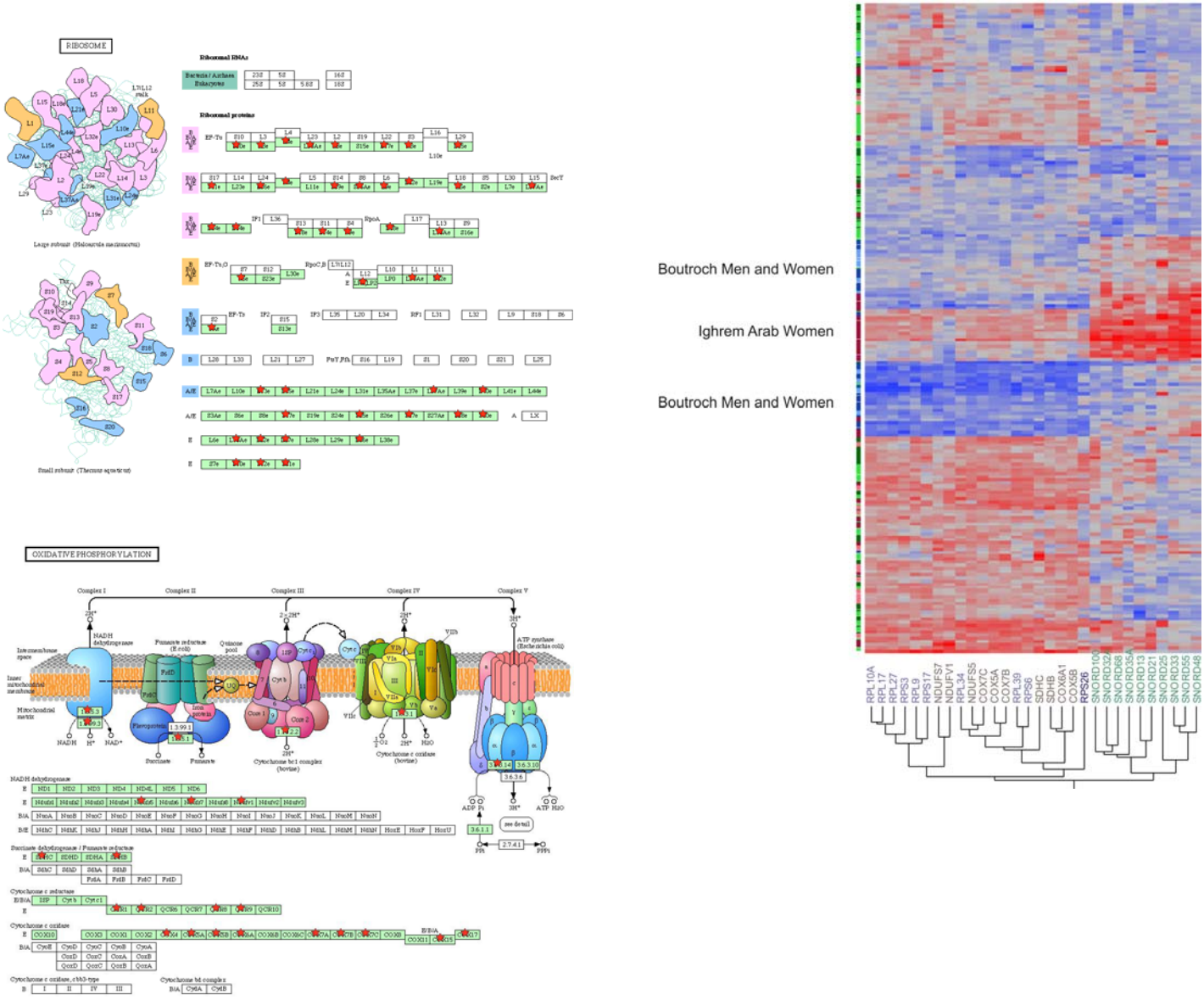
Supplementary Figure 3. Profiles of 4 representative genes showing relative fluorescence intensity by gender and location



Supplementary Figure 3. Profiles of 4 representative genes showing relative fluorescence intensity by gender and location. Individuals are plotted from left to right with males (lighter shade) and females within Anza, Dchiera, Ighrem and Boutroch. This representation clearly shows elevated expression of *CREBBP* and *SNORD55* in the rural villages, with the exception of Ighrem men. *MATR3* is high only in Ighrem women, while *CACNA1I* expression is slightly higher in the city.

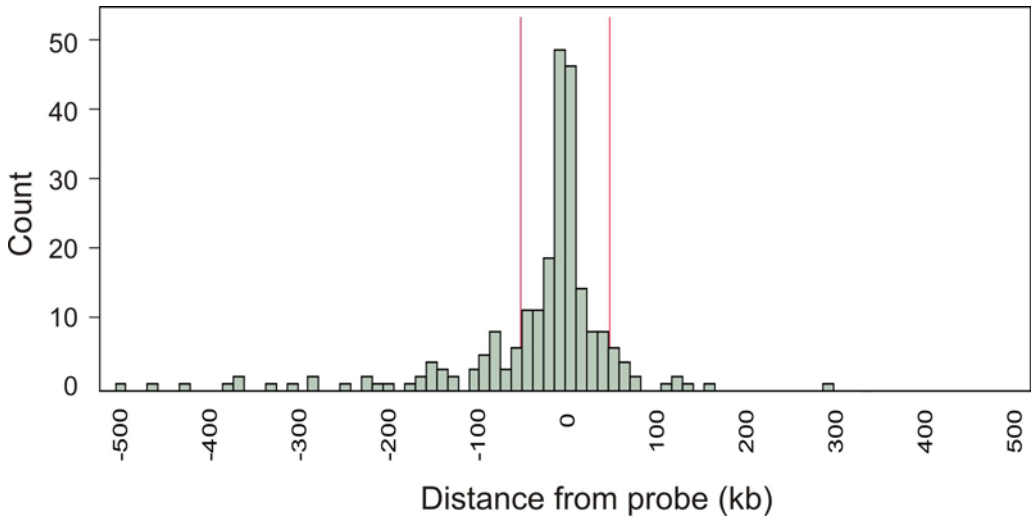
Supplementary Figure 4. Over-representation of ribosomal protein, SNORD and oxidative phosphorylation genes

a **b**



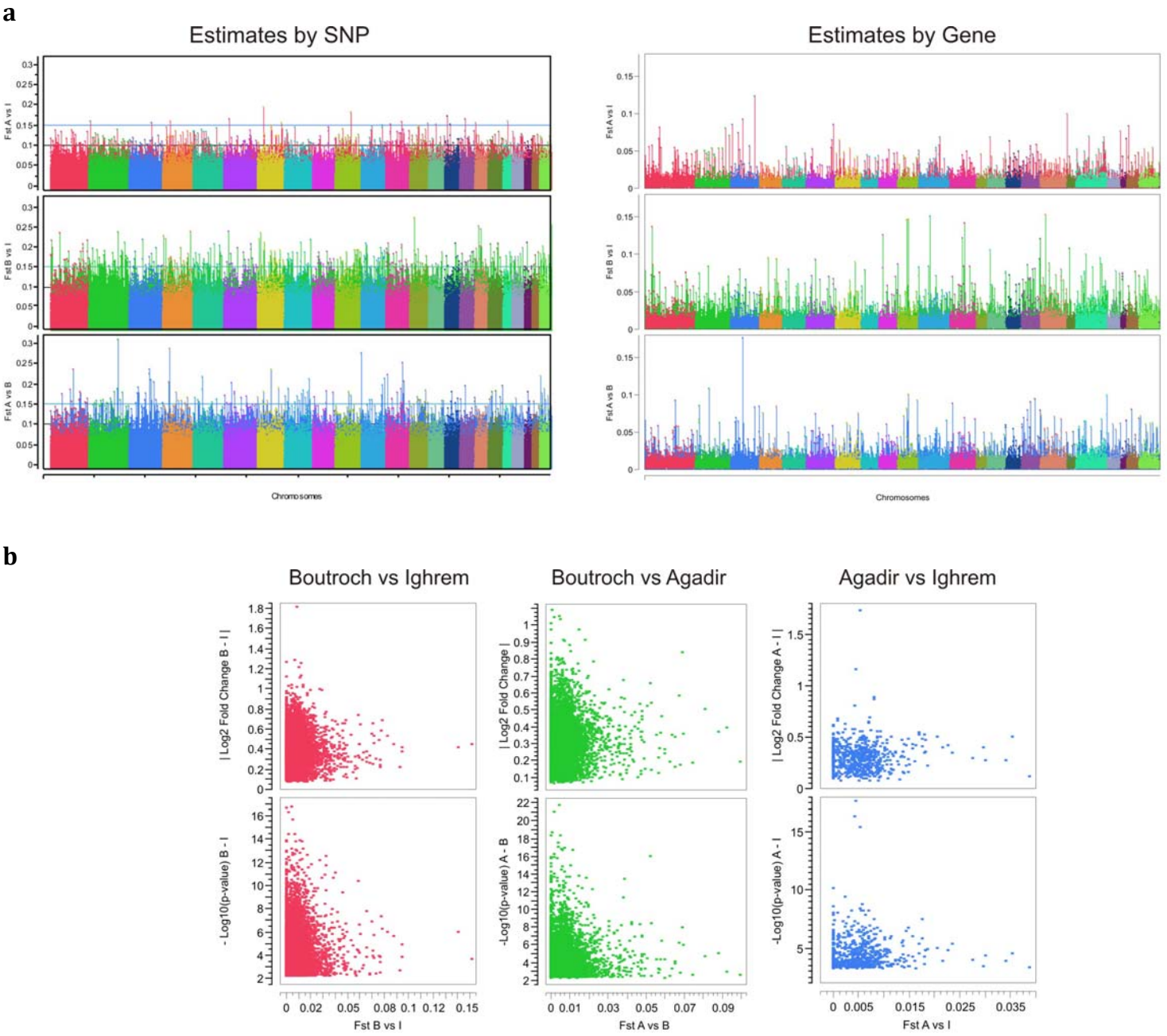
Supplementary Figure 4. **(a)** Over-representation of ribosomal protein (top) and oxidative phosphorylation (bottom) genes projected on KEGG representations of molecular complexes. Green genes are eukaryotic orthologs, and red stars indicate co-regulated genes between Boutroch and Agadir. **(b)** Two-way hierarchical cluster or representative gene networks, showing co-regulation of RP and OxPhos, but a distinct profile for SNORD genes. Each row is an individual, each column a gene, with red indicating relatively high transcript abundance, and blue low. Key to the left indicates whether the sample is from Agadir (green), Boutroch (blue) or Ighrem (red) with darker shading for women. Over half of the city residents and the Ighrem Arab women have a profile of high RP and OxPhos expression. By contrast, SNORD genes are highly up-regulated in the Ighrem Arab women, and relatively down-regulated in the Boutroch residents, implicating ethnic and gender differences by location that are most readily explained as cultural influences on gene expression.

Supplementary Figure 5. Distance of peak associations relative to the probe location



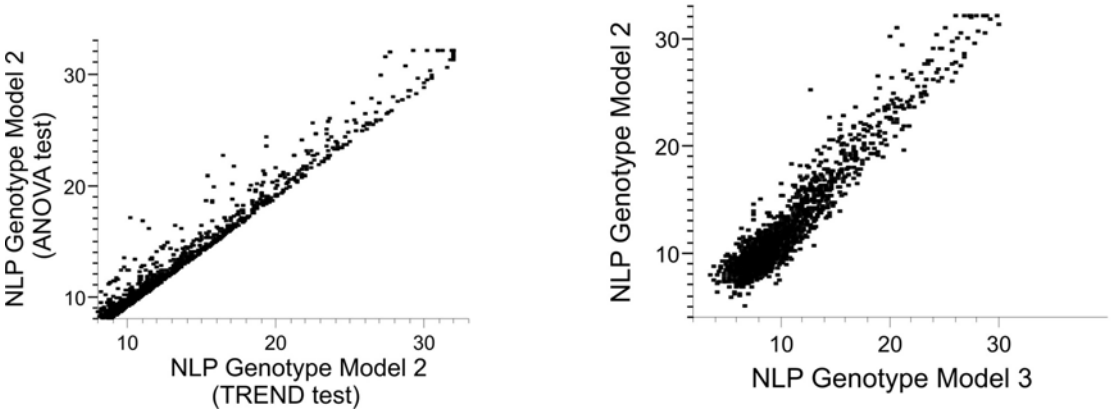
Supplementary Figure 5. Histogram of distance of peak associations relative to the probe location, for a sample of unambiguous *cis*-associations. The vast majority of associations mapping to the same chromosomal interval are within 50kb of the major transcription start site of the target gene as indicated by the red lines.

Supplementary Figure 6. Plots of genetic differentiation and relationship between genetic and gene expression divergence



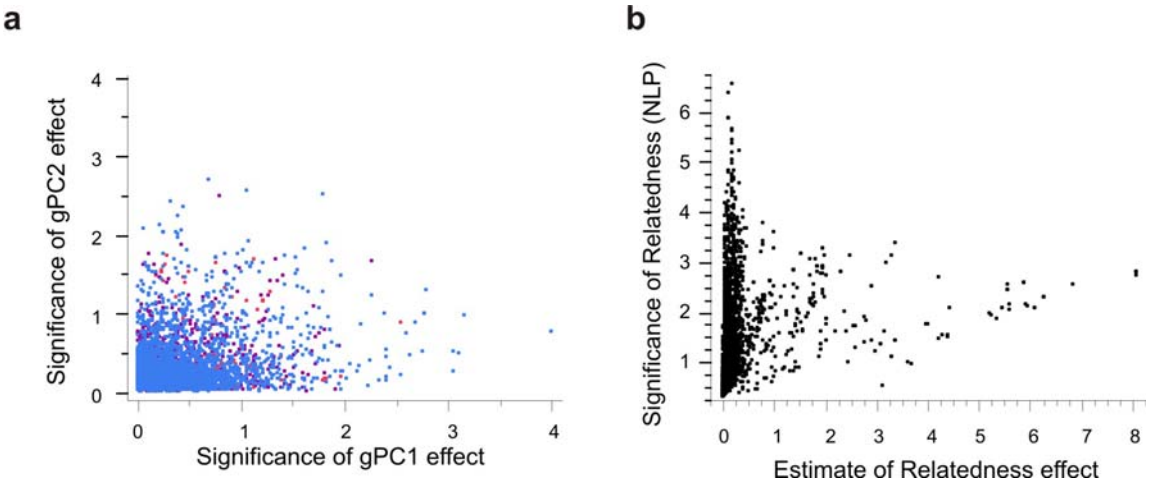
Supplementary Figure 6. **(a)** Plots of genetic differentiation (F_{ST} estimates) between pairs of populations for individual SNPs (left) and gene averages (right, mean of 22 SNPs per gene). SNPs are ordered by chromosome from 1 to 22 and the X at the right hand end. In each panel the comparisons are of Agadir and Ighrem, Boutroch and Ighrem, and Agadir and Boutroch from top to bottom. No SNPs exceed an F_{ST} of 0.3, and while a few hundred genes have elevated average F_{ST} values there is no strong divergence of chromosomes or chromosome intervals. **(b)** Plots showing absence of any global relationship between genetic and gene expression divergence. Top 3 plots show the relationship between fold difference in gene expression and F_{ST} for the comparisons of Boutroch and Ighrem, Boutroch and Agadir, and Agadir and Ighrem respectively. There is no tendency for high *p*-values to also show high gene expression divergence, or vice versa. Bottom three panels show similar comparisons of significance of expression divergence (negative log *p* values) against F_{ST}. Correlations in all six plots are non-significant ($P > 0.05$).

Supplementary Figure 7. Additional evidence for the robustness of associations



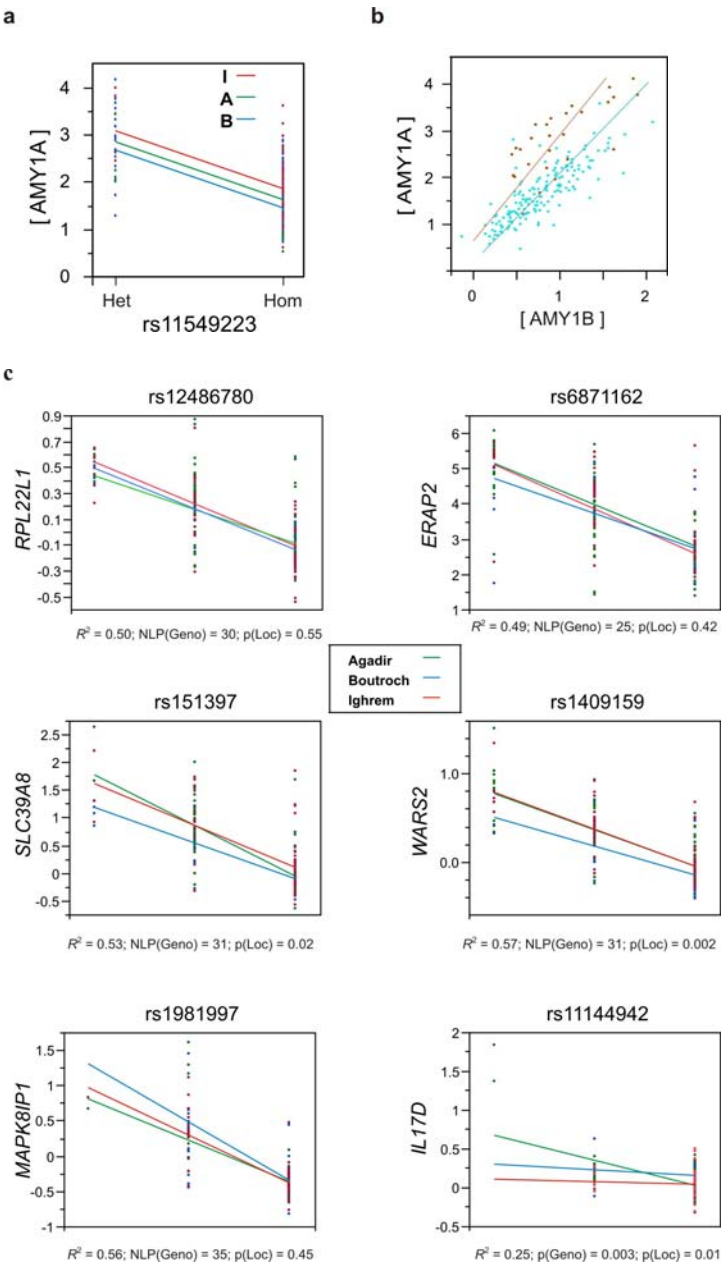
Supplementary Figure 7. Additional evidence for the robustness of associations described in this study is provided by plots of (**a**) the genotype versus allelic trend tests for each gene in Model 2 (which includes terms for Location, Gender and their Interaction), and (**b**) allelic trend tests for Model 2 against Model 3 (which includes terms for ethnicity, relatedness, and gender). Genotype tests are somewhat higher than allelic tests, but this is more likely due to small sample sizes of minor allele homozygotes in one or two of the populations that inflate the signal, than to non-additivity. The correlation between the Model 2 and Model 3 genotype allelic trend scores is highly significant with a correlation greater than 0.9.

Supplementary Figure 8. Additional evidence for the absence of genotype-by-environment interactions



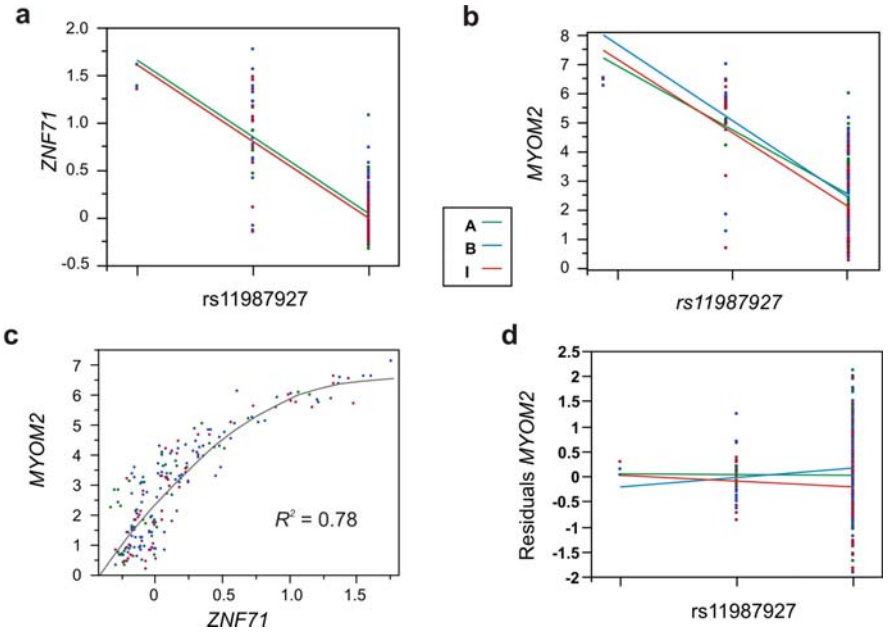
Supplementary Figure 8. Additional evidence for the absence of genotype-by-environment interactions is provided by these plots of (**a**) the significance of the effect of the genotypic estimates of ethnicity captured by PC1 (Sub-Saharan – North African) and PC2 (Arab-Amazigh), and (**b**) the significance against the magnitude of the relatedness effect. In (**a**) SNPs with minor allele frequency less than 0.05 are colored red, and less than 0.1 are purple: there is no global influence of allele frequency on the ethnicity effect. Note that the strongest significance values for the genotype effect are approximately 0.001, well above genome-wide significance. A few genes have more significant values for the relatedness term, but these do not survive multiple comparison correction either, and are not the SNPs with the largest estimated magnitude of effect.

Supplementary Figure 9. Examples of *cis* and *trans* associations



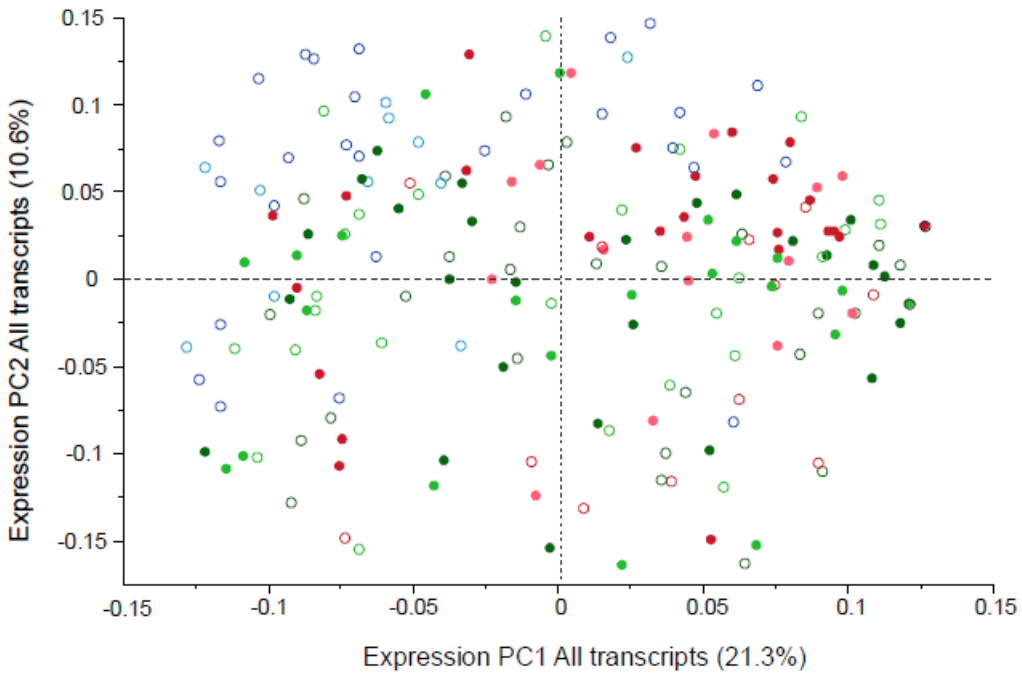
Supplementary Figure 9. A *trans*-eSNP in the *ATCG1* gene has a highly significant influence on expression of *AMY1A* **(a)**, whose expression is highly correlated with that of *AMY1B* **(b)**. However, there is no association between rs11549223 and *AMY1B* expression, and this plot shows that the combined effect of genetic and environmental influences that produce the correlation between the transcripts is constant across the two genotypes. Heterozygotes are brown, homozygotes blue-green in **(b)**, and the slope of the correlation is not significantly different, consistent with an additive effect of the SNP on *AMY1A* expression. **(c)** Six further examples of the robustness of genotype effects to location differences. Each plot shows the distribution of transcript abundance estimates for the three genotypes at the indicated eSNP, with green individuals from Agadir, blue from Boutroch, and red from Ighrem. In each case, the slope of the trend is almost identical for the three populations, and hence the Genotype-by-Location interaction is non-significant. The top 4 examples are *cis*-eSNPs, with no effect of Location for the top 2, and a small effect in the middle two. Typically, Boutroch is the location with the largest difference in expression. The bottom left example is a *trans* association. The bottom right is a cis-association that appears to show crossing of line means and a weak interaction effect, but this is solely due to two outlier individuals who both happen to be in Agadir; if the heterozygotes and major allele homozygotes are compared, there is no genotype effect at all.

Supplementary Figure 10. **Conditional dependence analysis of the *MYOM2* eSNP rs11987927**



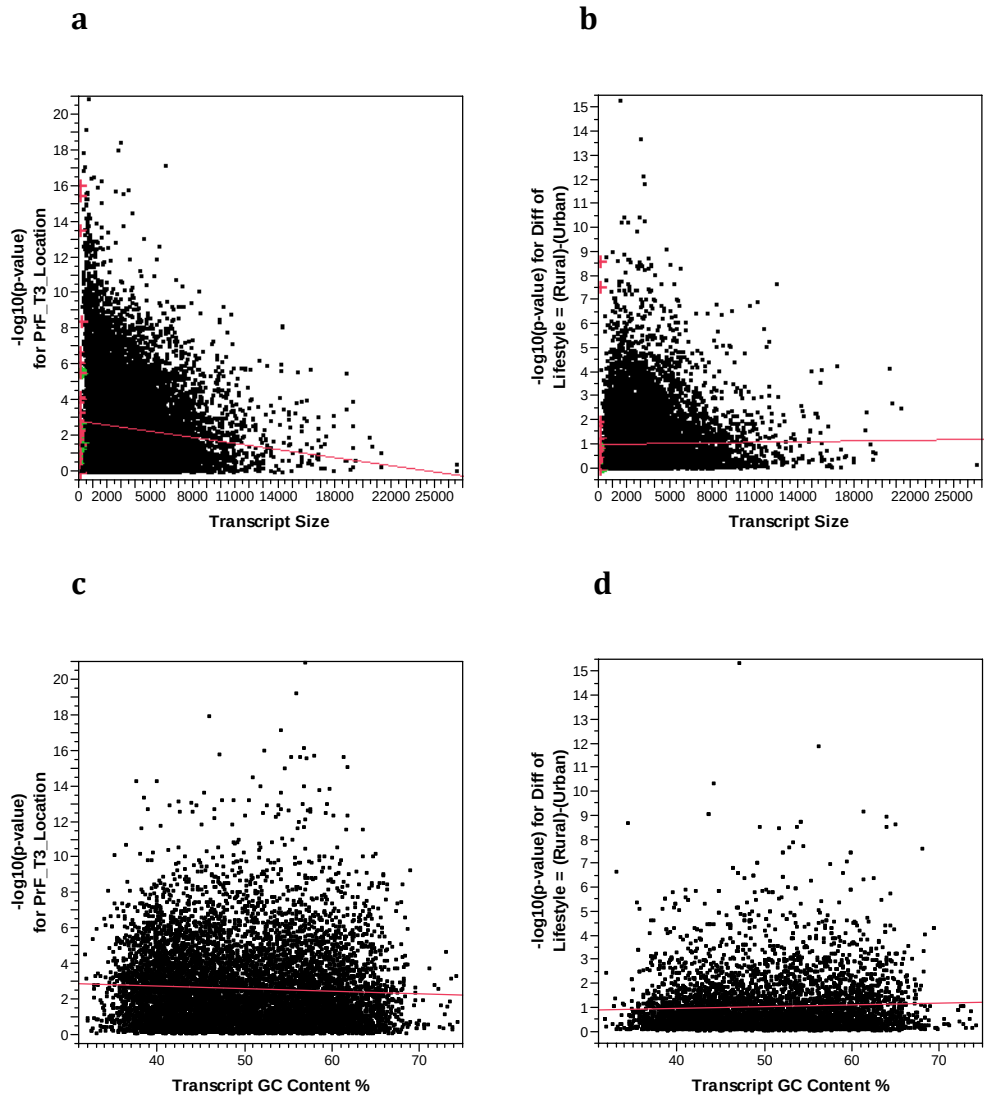
Supplementary Figure 10. Surprising conditional dependence analysis of the *MYOM2* eSNP rs11987927 on expression of *MYOM2* in *cis* (**a**: $p \sim 10^{-15}$) and *ZNF71* in *trans* (**b**: $p \sim 10^{-27}$). The two transcripts are highly correlated (**c**), and the *MYOM2* residuals from a quadratic regression of *MYOM2* on *ZNF71* are completely explained by the expression of *ZNF71* independent of the rs11987927 eSNP (**d**). The fit of *ZNF71* on *MYOM2* expression produces the same result, but with considerably less variance explained (not shown). The apparent ability of variation in expression of the target gene to explain variation in the abundance of the linked transcript is most likely due either to measurement error, as discussed in ref. 30, or covariance with another hidden variable. The less likely biological scenario is that rs11987927 is in LD with another variant that influences ZNF1 expression, which feeds back on MYOM2.

Supplementary Figure 11. **Principal components of the transcriptome variation**



Supplementary Figure 11 . Principal components of expression variation for the entire set of expressed transcripts. Boutroch individuals in blue, Ighrem in red, and Agadir in green; females darker shading; and self-reported Arabs as filled symbols and Amazigh as open circles. The plot indicates a similar overall profile to that shown in **Figure 4** for the most differentially expressed genes, but with less separation of the population groups.

Supplementary Figure 12. **Relationship between transcript size and GC content, and significance of differential expression**



Supplementary Figure 12. Absence of relationship between transcript size and significance of differential expression **(a)**. Plot of significance (NLP) against predicted RefSeq transcript size for the 3-way location effect (Agadir vs Boutroch vs Ighrem). While the slope appears to indicate higher significance for smaller transcripts, the R-squared value is just 0.013, and the relationship is driven by the low significance of transcripts above 10kb in length. **(b)** Similar plot for the Lifestyle effect (Rural versus Urban). In both plots, SNORD (red) and SNORA (green) transcripts are among the smallest detected on the array. **(c)** Absence of an effect of transcript GC content on significance of differential expression. NLP for the three-way location contrast against GC content for all transcripts, and **(d)** similarly for the lifestyle (rural versus urban) contrast.