

ORIGINAL ARTICLE

Demographic history and adaptation account for clock gene diversity in humans

I Dall'Ara, S Ghirotto, S Ingusci, G Bagarolo, C Bertolucci and G Barbujani

Circadian clocks give rise to daily oscillations in behavior and physiological functions that often anticipate upcoming environmental changes generated by the Earth rotation. In model organisms a relationship exists between several genes affecting the circadian rhythms and latitude. We investigated the allele distributions at 116 000 single-nucleotide polymorphisms (SNPs) of 25 human clock and clock-related genes from the 1000Genomes Project, and at a reference data set of putatively neutral polymorphisms. The global genetic structure at the clock genes did not differ from that observed at the reference data set. We then tested for evidence of local adaptation searching for F_{ST} outliers under both an island and a hierarchical model, and for significant association between allele frequencies and environmental variables by a Bayesian approach. A total of 230 SNPs in 23 genes, or 84 SNPs in 19 genes, depending on the significance thresholds chosen, showed signs of local adaptation, whereas a maximum of 190 SNPs in 23 genes had significant covariance with one or more environmental variables. Only two SNPs from two genes (*NPAS2* and *AANAT*) exhibit both elevated population differentiation and covariance with at least one environmental variable. We then checked whether the SNPs emerging from these analyses fall within a set of candidate SNPs associated with different chronotypes or sleep disorders. Correlation of five such SNPs with environmental variables supports a selective role of latitude or photoperiod, but certainly not a major one.

Heredity (2016) 117, 165–172; doi:10.1038/hdy.2016.39; published online 15 June 2016

INTRODUCTION

Many biological processes show circadian rhythms, that is, regular fluctuations along 24 h of the day, anticipating daily changes in the environment. A system of endogenous oscillators presents in almost all tissues and organs generates and coordinates such fluctuations, and it includes, in mammals, a central pacemaker located in the suprachiasmatic nuclei of the hypothalamus (Mohawk *et al.*, 2012). The main role of the circadian clock is to synchronize body circadian rhythms with those of the external environment, with light and temperature being the most important cyclic factors (or Zeitgebers) entraining the clock (Hut and Beersma, 2011).

At the molecular level, the circadian clock involves interlocked positive and negative transcriptional/translational feedback loops between clock genes and their protein products, together with a multilevel post-translational regulation of key clock components (Mohawk *et al.*, 2012). In mammals the positive loop is formed by the basic HLH-PAS proteins CLK (or its paralog NPAS2) and BMAL that heterodimerize and bind E-box elements in the regulatory regions of the *Period* (*Per*) and *Cryptochrome* (*Cry*) genes. The negative feedback loop is constituted by PER/CRY heterodimers that repress their own transcription by inhibiting CLK/BMAL activity. Moreover, the CLK/BMAL complex activates the transcription of two nuclear receptors, *Rev-Erba* and *Rora*, the protein products of which regulate *Bmal1* expression to bind retinoic acid-related orphan receptor response elements in the *Bmal1* promoter. Post-transcriptional modifications (such as phosphorylation, ubiquitination, sumoylation) are also crucial in the regulation of circadian protein turnover

(Partch *et al.*, 2014). For instance, mutations in the Casein kinase 1 ϵ and δ (*Ckl1e* and *Ckl1d*) alter the length of the circadian period in humans and mice (Xu *et al.*, 2005; Lee *et al.*, 2009).

Several mutations in human clock genes, such as *Per*, *Cry*, *Tim*, *Clk* and *Bmal1*, have been associated with abnormal circadian phenotypes and sleep–wake (see Allebrandt *et al.*, 2010; Hida *et al.*, 2014) but there is reason to believe that a greater number of serious and less serious disorders may be related with disfunctions of clock genes. Because the circadian clock is so tightly connected with the environment, it seems logical to expect that adaptation played a significant role in its evolution (Coop *et al.*, 2009). However, the (so far, limited) analyses of diversity of human clock genes have led to contradictory conclusions, with different studies supporting either a role of geographically variable positive selection (Cruciani *et al.*, 2008) or the simple effects of gene flow and genetic drift (Ciarleglio *et al.*, 2008) or both (Forni *et al.*, 2014).

Inferring adaptation is no trivial task (Li *et al.*, 2012). Only for a handful of human loci there is convincing evidence, from both genomic analyses and *in vitro* or *in vivo* studies of functional differences between genotypes, that some well-identified environmental factors shaped patterns of variation. Examples include skin pigmentation (Parra, 2007), the ability to digest lactose (Bersaglieri *et al.*, 2004) and the malaria-related polymorphisms at the Duffy blood group (Hamblin and Di Rienzo, 2000) and G6PD (Tishkoff *et al.*, 2001) loci. For many other loci, evidence suggestive of positive selection pressures does exist, but the patterns in the data are also consistent with neutral evolutionary mechanisms related with

demographic history (Harris and Meyer, 2006), that is, the processes leading to the anatomically modern human dispersal over the whole planet and successive local events.

In this study, we assembled a large database of DNA polymorphisms in human clock and clock-related genes, and used summary statistics to compare observed data with expectations generated under the null hypothesis of selective neutrality. We could identify loci in which populations sharply differ from each other. In some instances, these differences were detected between populations living at different latitudes, and hence going through different modes of seasonal fluctuation of environmental variables, such as the photoperiod. This way, we could identify several single-nucleotide polymorphisms (SNPs) showing evidence of positive selection, namely local adaptation, clinal adaptation and, in two cases, both.

MATERIALS AND METHODS

Genomic data

We collected genetic data for SNPs genotyped in 2504 individuals, belonging to 26 worldwide populations from the 1000Genome browser (<http://www.1000genomes.org/>; 1000 Genomes Project Consortium *et al.*, 2015). Among the populations considered, 7 are from Africa, 5 from Europe, 5 from South Asia, 3 from East Asia and 4 are defined as Admixed American (Figure 1). We assembled three genetic data sets: two consisting of clock gene polymorphisms and one with presumably neutral SNPs. The first includes all SNPs in 25 autosomal genes, namely clock, clock-controlled and clock-related genes (Supplementary Table S1). Out of the set of 135 162 SNPs we retained a subset of polymorphisms showing a level of pairwise linkage disequilibrium lower than $r^2 = 0.3$ (as estimated in the entire data set) by the PLINK v1.07 tool (Purcell *et al.*, 2007), using the following command: `plink -file data -indep-pairwise 50 5 0.3`. This pruned subset contains 116 440 SNPs and we called it Candidate data set.

Among these SNPs, we identified in the literature 15 polymorphisms belonging to 11 clock genes (Supplementary Table S2), for which an association with human sleep disorders was reported, such as the advanced sleep phase syndrome and the delayed sleep phase syndrome, or with changes in circadian phenotypes, at least in one world population. The data set including phenotypic information, here referred to as the 'Phenotype-associated' data set, is not

independent from the Candidate data set, but no analysis was run in parallel on both. Finally, we created a 'Reference' data set using the NRE software (<http://nre.cb.bscb.cornell.edu/nre/>; Arbiza *et al.*, 2012) applying Patin *et al.* (2009) criteria. It consists of 200 000 autosomal neutral sites, sampled at random in noncoding regions and at least 200 kb away from any known or predicted gene.

Environmental data

We included in the analyses seven environmental factors, that is, photoperiod, humidity, radiation fluxes, precipitation, temperature and two proxy variables summarizing climatic diversity: latitude and longitude. We also considered the geographic distance from an arbitrary point in East Africa (Addis Ababa) roughly representing the origin of the expansion processes of anatomically modern humans, thus taking into account one crucial demographic event affecting the allele frequency distribution.

Geographic distances from Addis Ababa were calculated as great circle distances. In agreement with previous studies (Ramachandran *et al.*, 2005), to better reflect in those distances the likely routes of human dispersal between continents, distances between populations of different continents were calculated through five obligatory waypoints (across the Suez Canal, between Anatolia and Europe, in Northeastern Russia, Cambodia and Northwestern Canada).

To test for any effects of the seasonal variation of the photoperiod, we calculated for each population the difference in photoperiod between the Summer and the Winter solstices (http://aa.usno.navy.mil/data/docs/RS_One-Year.php).

The NCEP/NCAR (National Centers for Environmental Prediction/National Center for Atmospheric Research) Reanalysis Project at the NOAA/ESRL (National Oceanic and Atmospheric Administration/Earth System Research Laboratory) Physical Sciences Division (<http://www.esrl.noaa.gov/psd/data/reanalysis/reanalysis.shtml>) contains raw data of humidity and solar radiation fluxes starting from 1 January 1948 until 1 January 2015. For each population, the mean relative humidity was calculated as the ratio between the actual vapor density and the saturation vapor density. Regarding the solar radiation fluxes, we took into consideration the net shortwave and longwave radiation fluxes. The former represents a measure of the difference between the incoming solar shortwave radiation and the outgoing shortwave radiation from the earth surface, whereas the latter is a measure of the difference between the outgoing longwave radiation from the earth surface and the incident atmospheric

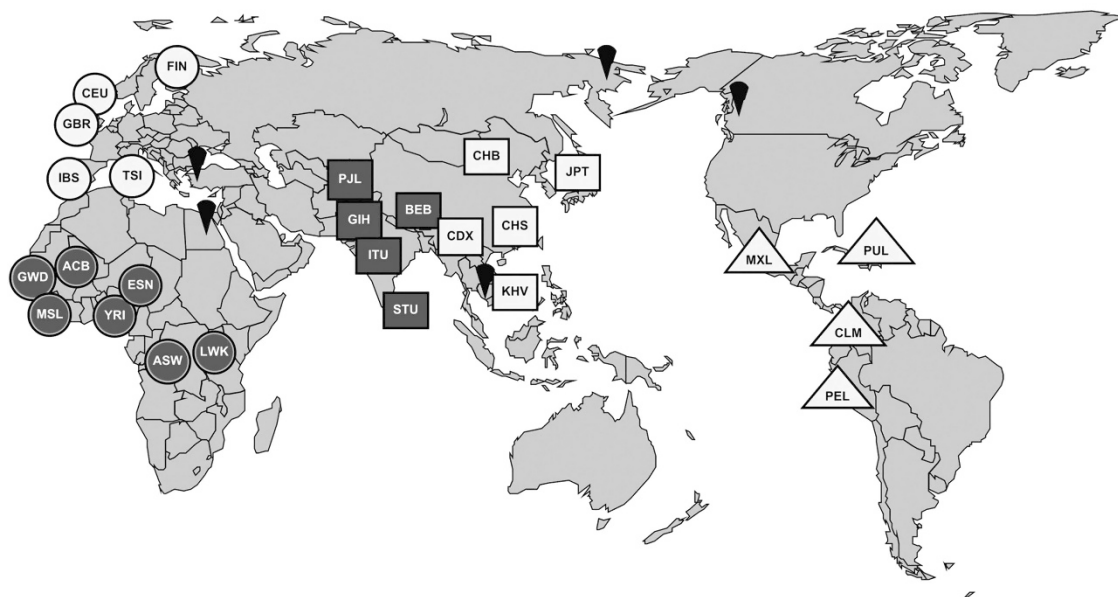


Figure 1 Worldwide distribution of 26 populations, belonging to 5 different groups: African (full circle), European (empty circle), South Asian (full square), East Asian (empty square) and Admixed American (empty triangle). Black triangles represent the five waypoints used to estimate geographic distances between pairs of populations (Ramachandran *et al.*, 2005).

longwave counter-radiation. For each sample, in reference to the period 1 January 1948–1 January 2015, we calculated the ratio between the average of net longwave and of net shortwave radiation fluxes.

We obtained information for both the temperature and the precipitation rate from the Climatic Research Unit database (<http://www.cru.uea.ac.uk/>). Monthly average temperatures were measured starting from 1850, in stations at different elevations, often using different methods. In order to avoid biases that could result from these problems, temperatures in the database were reduced to anomalies from the period with best coverage, 1961–1990. We calculated, for each population, the mean annual temperature, considering the number of years for which data are available. Conversely, data about precipitations 1900–1998 are available for all 26 populations investigated in this study. For each of them, we calculated the mean precipitation rate, expressed in tenths of a millimetre.

Furthermore, we included in our analysis the longitude and latitude of each population. The latter is a common proxy for other, unspecified environmental variables that may have an effect of climate and/or photoperiod (Gunther and Coop, 2013).

Population structure

As a preliminary step, we performed a principal component analysis (PCA) using the `snpGdsPCA` function of the `SNPRelate` R package (Huber *et al.*, 2015).

Then we moved to a model-based population structure analysis on the Candidate and Reference data set using the `ADMIXTURE` software (<http://www.genetics.ucla.edu/software/admixture/>; Alexander *et al.*, 2009). To identify the most supported number (K) of ancestral populations we calculated the cross-validation error considering the number of K from 1 to 20. In both cases five ancestral populations seem to best explain the genetic variation contained in the data. We then clustered individual genotypes exploring a range of K -values between 2 and 12, and repeating the analysis 50 times for each K -value. The results were finally summarized and plotted using the `CLUMPP` (<http://www.stanford.edu/group/rosenberglab/clumpp.html>; Jakobsson and Rosenberg, 2007) and `Distruct` software packages (<http://www.stanford.edu/group/rosenberglab/distruct.html>; Rosenberg *et al.*, 2002).

Testing for the effects of natural selection

Positive selection leaves a signature in the genome that we tried to detect in three ways. First, we used `BayeScan` (<http://cmpg.unibe.ch/software/BayeScan/>; Foll and Gaggiotti, 2008), a differentiation-based method assuming a Bayesian framework in which F_{ST} is a model parameter. For each locus, and taking advantage of a reversible jump Markov Chain Monte Carlo algorithm, `BayeScan` estimates the scale of evidence in favor of a model including selection versus a model without selection. In particular, `BayeScan` scans for local adaptation in allele frequency data modeling separately a coefficient (β) assuming an island model of demography and thus regarded as population specific, and a locus-specific coefficient (α) that reflects the strength of selection acting on a particular locus. For each SNP, `BayeScan` estimates the posterior distribution under neutrality ($\alpha=0$) and allowing for selection ($\alpha\neq 0$), thus computing the posterior odds ratio (PO) as an estimate of the support for the model of local adaptation compared with neutral demography. This way, if α is positive, the locus considered has potentially undergone positive selection, whereas if α is negative, the locus is considered subjected to balancing selection. Outlier detection levels depend on the estimation of the PO that can be evaluated through the Jeffrey's scale of evidence (Jeffreys, 1961). To control against false positive results we calculated the expected value of $\log_{10}PO$ yielding a 1% of false discovery rate. We performed the `BayeScan` analyses on both the Candidate and the Reference data sets under identical run conditions, consisting of prior odds of 1000, meaning that we considered the selection model 1/1000 as likely as the neutral model for any given SNP, 20 pilot runs, 50 000 burn-in iterations followed by 50 000 output iterations with a thinning interval of 10, resulting in 5000 iterations for posterior estimation. As suggested in the `BayeScan` manual, we removed from the analysis loci with very low minor allele frequencies (<0.005), so that 14 707 SNPs of the Candidate data set and 27 626 SNPs of the Reference data set were in fact analyzed. We defined two significance thresholds. The first one was a Model-based significance threshold, namely the expected value of the \log_{10} of the PO (the support for

the model of local adaptation relative to neutral demography) that would yield a 1% false discovery rate based on the reference SNPs. The second threshold was empirical, hence called Reference-based significance threshold, and it was simply the 1% upper confidence limit of the F_{ST} distribution in the Reference data set; all SNPs in the Candidate data set showing F_{ST} values beyond this limit were considered significant.

If the population structure does not correspond to Wright's island model, `BayeScan` may give biased results (Excoffier *et al.*, 2009). To circumvent this problem we turned to the second method, based on a hierarchical island model (Excoffier *et al.*, 2009) and implemented in the `Arlequin` software (v3.5.1.2; Excoffier and Lischer, 2010) that specifically takes into account coancestry among related subpopulations. Accordingly, we generated the expected neutral F_{ST} distribution conditioned on heterozygosity using a subsample of 1000 reference SNPs. To model the hierarchical genetic structure, in agreement with Excoffier *et al.* (2009), we defined 10 groups, each consisting of 100 demes. The expected distribution of F_{ST} values for the reference SNPs was then generated simulating the effects of this structure for 50 000 replicates by the `Arlequin` software. The same approach was then used to calculate F_{ST} on the Reference data set. Again, in agreement with the previous (`BayeScan`) analysis of local adaptation, we defined a Reference-based threshold corresponding to the candidate SNPs to an F_{ST} value greater than the 99% upper limit of the reference F_{ST} distribution, and a Model-based threshold related to the simulated F_{ST} distribution based on 50 000 simulations.

In our third approach, we looked for local selection reflecting an environmental gradient. One can detect cases in which the covariance between allele frequencies and ecological variants exceeds the expected covariance at neutral loci by the method proposed by Coop *et al.* (2010) implemented in the program `BAYENV2` (<http://gcbias.org/bayenv/>). First, we computed the neutral covariance matrix based on 10 000 reference SNPs, thus summarizing the pattern of allelic frequency variance among the 26 populations according to a simple drift model. This matrix of population differences (Ω) was then used to control for population history when testing for covariance between environmental variables and the population-specific allele frequencies at a given SNP. For each locus `BAYENV2` computes the ratio of the posterior probability (PO) of the model of adaptation vs random drift; the PO and the associated Bayes factor (BF) represent the support for the model of local adaptation with respect to a model of random drift. This parametric method assumes a linear effect of the environmental variable, and this can lead to spurious correlations in the presence of strong outliers. `BAYENV2` also allows to estimate the nonparametric Spearman's ρ statistic that is less affected by extreme values. We analyzed each SNP individually and determined the distribution of PO, BF and ρ separately for reference and candidate SNPs. In order to identify candidate polymorphisms really showing a strong departure from null expectations, we defined two significance thresholds. Hereafter, we shall speak of Single threshold when, for a given SNP, the BF of the model of adaptation vs the null model is >10 ; combined threshold means instead that a SNP has both BF >10 and a ρ -value >0.25 , as in Jaramillo-Correa *et al.* (2015).

RESULTS

Analysis of population structure

We started from exploratory analyses of two genomic data sets: the Candidate and the Reference data sets.

The PCA plots of the Reference and Candidate SNPs show remarkable similarities (Figures 2a and b). In both cases, the African populations are spread along the first PC axis (accounting in both cases for $<1.5\%$ of the total variance); admixed American populations occupy intermediate positions between Africa and, to the left, the Eurasian populations that are distributed along the second PC axis (accounting in both cases for $<0.5\%$ of the total variance). At this level of resolution, it is impossible to judge whether the small differences observed between data sets are suggestive of specific selective processes affecting loci of the Candidate data set; we can only say that these differences, if any, are not obvious.

We then looked for the best clustering of genotypes of the Reference and Candidate data sets in a number of clusters from $K=2$ to $K=12$.

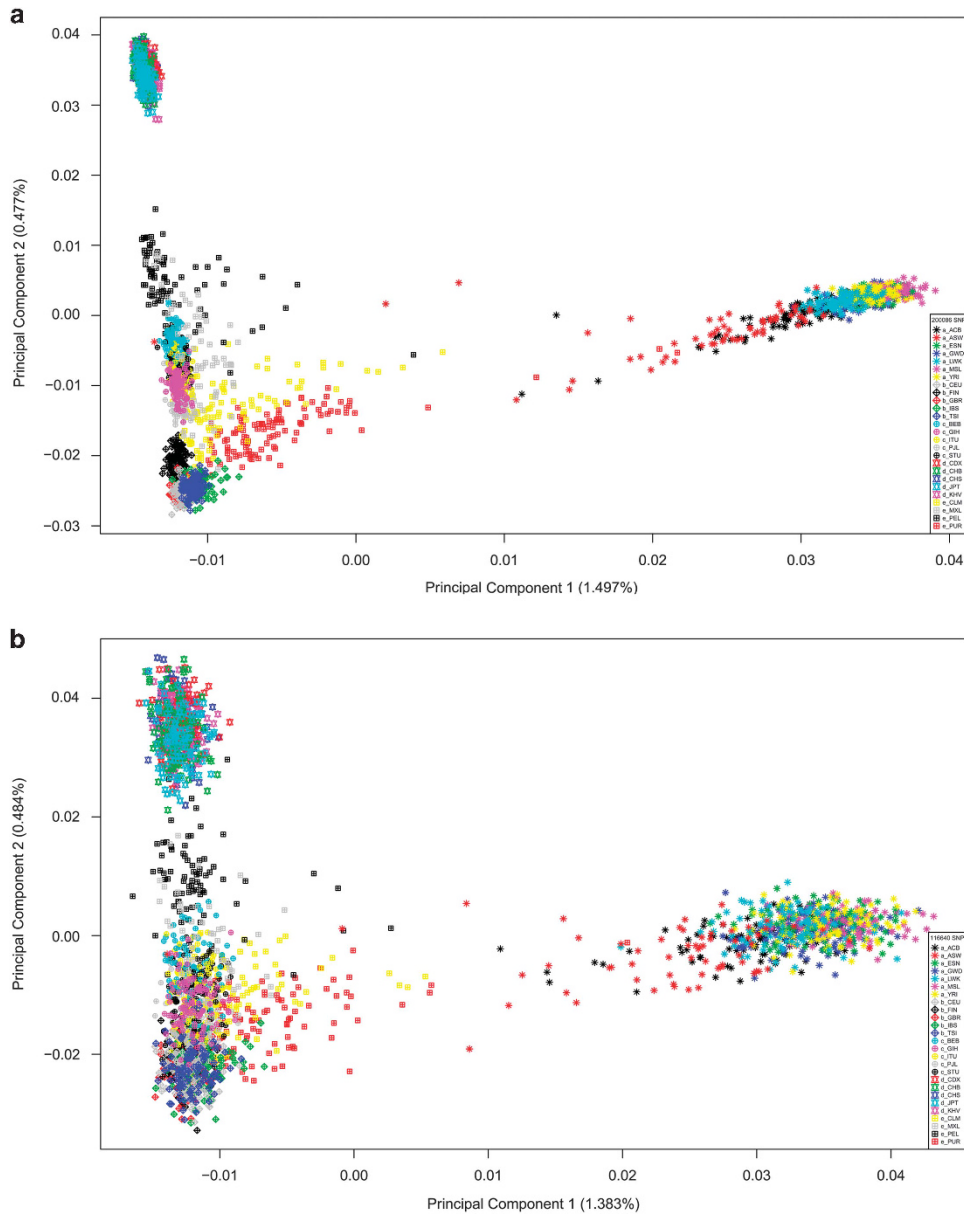


Figure 2 Plots of PCA inferred from the Reference (a) and Candidate (b) data sets.

In agreement with previous studies (Rosenberg *et al.*, 2002), at $K = 3$, clusters correspond to the main continents (orange for Africa, blue for Europe and yellow for Asia), with the American populations showing, once again, evidence of admixture on a recognizable genomic background. The most likely number of clusters (minimum cross-validation error) was $K = 5$ (Supplementary Figures S1A and B). A remarkably similar pattern was observed for the Reference data set (Supplementary Figure S1A) that also showed the minimum cross-validation error at $K = 5$. In short, these explorative analyses did not suggest that the distribution of SNPs involved in the human circadian machinery is different from the distribution of random autosomal SNPs.

Testing for local adaptation

BayeScan estimations of average population divergence were slightly higher for candidate SNPs than for the reference SNPs (0.125 vs 0.117). Furthermore, the number of SNPs that exceeded the 1%

Model-based significance threshold is higher than the number of SNPs showing evidence of local adaptation using the Reference-based 1% threshold (Table 1). Support for selection on candidate SNPs was very strong, with 74% of SNPs having $\log_{10}PO$ ratios > 3 . This value, according to the Jeffrey’s scale of evidence, is considered to be a *Decisive* support for the selection model (Jeffreys, 1961). Consistent with this interpretation, the estimated locus-specific effect (α) was higher in the Candidate than in the Reference data set (0.32 vs 0.30, on average), reaching a mean value of 1.4 in the 1250 positive selected polymorphisms of the Candidate data set. The positively selected polymorphisms under both thresholds are reported in Supplementary Table S3.

Population divergence estimated through the hierarchical model identified 91 SNPs in 22 genes having extreme F_{ST} values (> 0.3 , Supplementary Table S3). As it is shown in Figure 3, this is a conservative choice, as 0.3 greatly exceed the 99% confidence interval of the values simulated under a neutral model, represented by the

Analyses of the ‘Phenotype-associated’ data set

Finally, we wanted to understand whether SNPs in the ‘Phenotype-associated’ data set have been possible targets of adaptation phenomena. None of our Phenotype-Associated SNPs showed significant departures from null expectations in the previous analyses based on extreme F_{ST} values (that is, BayeScan and Hierarchical Model). However, when considering candidate SNPs with $BF > 99\%$ of the reference values in the BAYENV2 analysis, we could identify 5 polymorphisms showing significant correlation with at least one of the environmental factors analyzed here. These polymorphisms, falling in five crucial clock genes, along with the correlated environmental variables, are detailed in Table 5.

Table 2 Genes showing at least 10% of single-nucleotide polymorphisms (SNPs) with evidence of local adaptation

Reference based		Model based	
Gene	% Of SNPs	Gene	% Of SNPs
<i>RORA</i>	18	<i>RORA</i>	25
<i>NPAS2</i>	11	<i>CK1b</i>	19
<i>BMAL1</i>	10	<i>AVPR1B</i>	12
		<i>TIM</i>	11

Table 3 Number of single-nucleotide polymorphisms (SNPs) showing evidence of correlation with environmental variables, using both the ‘Single’ and ‘Combined’ thresholds in the BAYENV analysis

Significant threshold	Single	Combined
No. of SNPs	190	69
No. of genes	23	20

Table 4 Two single-nucleotide polymorphisms (SNPs) showing both evidence of local adaptation and significant correlation with, at least, one environmental factor

Significant threshold	Gene	SNP	Correlated environmental variables
Reference based	<i>NPAS2</i>	<i>rs72976842</i>	Latitude Ratio of L/S wave radiation fluxes
Model based	<i>AANAT</i>	<i>rs4647868</i>	Ratio of L/S wave radiation fluxes

Abbreviations: L/S, long/short; SNP, single-nucleotide polymorphism. For each SNP the significance threshold and the correlated environmental variable is reported.

Table 5 Five SNPs in the phenotype-associated data set showing significant correlation with at least one environmental variable

Gene	SNP	Correlated environmental variable(s)	Chronotypes associated	References
<i>NPAS2</i>	<i>rs7598826</i>	Distance from Addis Ababa Ratio of L/S wave radiation fluxes Mean longwave radiation flux	Alteration of sleep duration	Allebrandt <i>et al.</i> (2010)
<i>PER2</i>	<i>rs934945</i>	Mean humidity rate	Diurnal preferences	Lee <i>et al.</i> (2011)
<i>PER3</i>	<i>rs228730</i>	Mean precipitation rate	DSPS	Archer <i>et al.</i> (2010)
<i>CK1d</i>	<i>rs7209167</i>	Photoperiod	Alteration of sleep duration	Allebrandt <i>et al.</i> (2010)
<i>OPN4</i>	<i>rs2675703</i>	Distance from Addis Ababa Longitude	Variation in sleep onset and chronotypes	Roekelin <i>et al.</i> (2012)

Abbreviations: DSPS, delayed sleep phase syndrome; L/S, long/short; SNP, single-nucleotide polymorphism. For each polymorphism, we reported the abnormal chronotypes or the sleep disorder to which the polymorphism has been previously associated.

DISCUSSION

Taken together, the results of our analyses do show support for the view whereby various components of the human circadian clock evolved under the effects of positive selection, but also that selective pressures were rather weak upon individual loci.

Previous studies in human populations investigated the association of the circadian clock genes with pathological conditions, such as metabolic and affective disorders. For example, Forni *et al.* (2014) analyzed whole genome polymorphism data in populations of the CEPH panel to investigate how the seasonal photoperiod variation might have exerted a selective pressure during the expansion of modern humans out of Africa. They suggested that this expansion influenced adaptive evolution at circadian regulatory loci and at risk variants for psychiatric and neurologic diseases. Lane *et al.* (2016) performed a genome-wide association study of self-reported chronotypes of > 100 000 UK individuals, identifying 12 variants in genes involved in the circadian clock mechanism and in the light-sensing pathways (possibly related to schizophrenia, educational attainment and body mass index). A similar study has been conducted by Hu *et al.* (2016), who searched for correlations between whole genome variants and a self-reported morningness in the 23andMe participant cohort, finding significant signals in circadian and phototransduction pathways. The present survey thus represents a broader effort to explore specifically human clock gene variation at the worldwide level and using polymorphism data coming from complete genomes; this way, we could draw inferences on their evolution and on the effect of potentially relevant factors in the environment. The Candidate- and the Phenotype-associated data sets we analyzed include SNPs associated either generally with the circadian clock or specifically with chronotype alterations.

Our initial analyses of population structure show that the distribution of the SNPs of the Candidate data set, including clock, clock-controlled and clock-related genes, does not depart from that observed in the analysis of neutral genome diversity (see also Rosenberg *et al.*, 2002; Li *et al.*, 2008). Although such a result does not mean that selection was irrelevant, certainly it does not point to simple selective mechanisms affecting single genes as major determinants of population differences.

Inferring positive selection from genomic data is complicated by the fact that patterns in the data may indeed reflect selection, or also the effects of population history, or a combination of both. For some 60 years now, the presence of outliers in the distribution of genetic variances, that is, of loci showing either very low or very high levels of variation, has been taken as suggestive of phenomena affecting single loci (such as selection), as opposed to the whole genome (such as genetic drift and gene flow) (Cavalli-Sforza, 1966). The availability of large amounts of information about nongenic regions

has increased the power of tests exploiting this basic idea (Novembre and Di Rienzo, 2009).

Because the Admixture and PCA analyses did not point to obvious differences between clock genes and presumably neutral genome regions, we chose to combine different approaches to detect the signature of selection at individual loci. Whether or not the island model assumed by BayeScan can faithfully represent migrational relationships at this scale is open to discussion. In principle, a hierarchical island model seems to be more realistic and more conservative than a standard island model, but, in practice, there is no way to tell. We thus decided to combine the two approaches in order to identify a (smaller) subset of SNPs showing significant departures from null expectations under both methods, and hence to have a higher chance to tell natural selection from the background of demographic history.

The results of our analyses point to several SNPs departing from null expectations in their distribution, some of them very significantly. Most of the polymorphisms with evidences of local adaptation were in the *BMAL1*, *NPAS2*, *TIM*, *RORA* and *CK1δ* genes, suggesting selection action might have affected all circadian clocks components (positive, negative and clock-related genes).

In addition, and contrary to what was observed in one (Cruciani *et al.*, 2008) but not all (Forni *et al.*, 2014) previous studies on human clock genes, we also found some evidence of correlation between circadian clock SNPs and environmental factors.

Two SNPs emerge from the joint analyses of local and clinal adaptation, namely *rs72976842* in *NPAS2* and *rs4647868* in *AANAT*. *NPAS2* (neuronal PAS domain protein 2) is a transcription factor paralog of *CLOCK* expressed mainly in the mammalian forebrain. It has been shown in mice that both clock proteins can independently heterodimerize with *BMAL1* in the suprachiasmatic nuclei to maintain molecular and behavioral rhythmicity (Reick *et al.*, 2001). An overlapping role of *CLOCK* and *NPAS2* has also been shown in the circadian expression of liver *FVII*, a key liver protease (Bertolucci *et al.*, 2008). In humans, epidemiologic evidence linked *NPAS2* with a variety of disorders, including winter depression (Partonen *et al.*, 2007). Furthermore, a *NPAS2* missense mutation (*rs2305160*) is associated with risk of cancer (Zhu *et al.*, 2008), but that SNP is not in linkage disequilibrium with the one we found to be positively selected, namely *rs72976842* ($r^2 = 0.014$).

The second locus, *AANAT* (*arylalkylamine N-acetyltransferase*), codes for a key enzyme involved in daily melatonin synthesis and its transcription is regulated by the circadian clock via the E-box promoter elements.

Evidence of correlation between ecological factors and five SNPs from the Phenotype-associated data set suggest that the genetic variation associated with some crucial components of the circadian clock is variously influenced by environmental factors (Table 5). Three such polymorphisms are in clock genes (*rs7598826* in *NPAS2*, *rs934945* in *Per2* and *rs228730* in *Per3*) and two in clock-related genes *CK1δ* (*rs7209167*) and *OPN4* (*rs2675703*).

We have already mentioned *NPAS2*; its SNP *rs7598826* is known to correlate with sleep duration (Allebrandt *et al.*, 2010). SNPs in *PERs* correlate with humidity and precipitation mean rates and have been associated with diurnal preferences (*Per2*, *rs934945*: Lee *et al.*, 2011; *Per3*, *rs228697*: Hida *et al.*, 2014) and delayed sleep phase syndrome (*Per3*, *rs228730*: Archer *et al.*, 2010). Both *Per2* and *Per3* participate in timekeeping mechanism in the central (suprachiasmatic nuclei) and peripheral (pituitary gland and lung) oscillators, respectively (Pendergast *et al.*, 2010). A recent study demonstrated that the phosphorylation of the *PER2* protein is the mechanism underpinning the 'temperature compensation', a key feature of the circadian clock

that allows the organisms to maintain a 24-h period regardless of the environmental temperature (Zhou *et al.*, 2015).

Per3 was deeply studied because some 10% of the human population is homozygous for the five-repeat allele of a VNTR (variable number tandem repeat) polymorphism within it, and these people show morning preferences (Archer *et al.*, 2010). In contrast, the homozygous genotype for the four-repeat allele has been associated with evening preferences (Archer *et al.*, 2010). We did not investigate this VNTR polymorphism because the standard algorithms we used to identify selection are not suitable for tandem repeat loci.

As for the *CK1δ* gene, it controls nuclear transport and degradation of core elements of the circadian clock such as *PER1* and *PER2* and determines the circadian period length through the regulation of the speed and rhythmicity of their phosphorylation (Partch *et al.*, 2014). It is intriguing to note that *rs7209167* in *CK1δ* shows a significant correlation with photoperiod and has been associated with the alteration of sleep duration (Allebrandt *et al.*, 2010). *CK1δ* is currently the target of pharmacological studies aimed at shifting or resetting the phase of circadian rhythm for the treatment of circadian rhythm disorder, such as the familial advanced sleep-phase syndrome determined by a missense mutation (T44A, *rs104894561*) (Xu *et al.*, 2005). *rs104894561* was not included in our analysis because population data are not available in 1000Genome browser.

Finally, *OPN4*, also called melanopsin, is a circadian photopigment expressed within the ganglion and amacrine cell layers of the mammalian retina (Lucas, 2013), mediating non-image-forming responses to light, such as the sleep induction, the pupillary light response and others. The mouse melanopsin gene is present in two fully functional isoforms: the short (*OPN4S*) and the long (*OPN4L*) isoform. Recent papers showed that these different isoforms of *OPN4* mediate different behavioral responses to light (Jagannath *et al.*, 2015). Both isoforms are also present in the human genome, but the functional results requires confirmation. *rs2675703* in *OPN4*, the SNP that we found correlated with two geographical parameters, has been previously associated with variation in sleep onset and chronotypes (Roekelin *et al.*, 2012). Recently, activator or inhibitor of melanopsin have been proposed as treatments to mimic light and darkness to manage sleep disturbances and mood disorder in normal and blind patients (Hatori and Panda, 2010).

Although all these findings make biological sense, and may have implications for the development of drugs against sleep disturbances and disorders, very few SNPs of this study correlate with some of the main environmental variables. We interpret this finding as suggestive that mechanisms underpinning the evolution of human circadian clock, much like those related with other complex traits, are the result of more complex interactions, ones that can hardly be discovered focusing only on SNP frequencies.

Investigations on how circadian systems are adapted to different latitudes in humans are at an early stage. Evidence of genetic differences likely because of selection caused by environmental conditions along latitudinal clines are mainly available for insects and fishes (Kyriacou *et al.*, 2008; O'Malley and Banks, 2008). Analysis conducted on *Drosophila melanogaster* (Kyriacou *et al.*, 2008) showed that latitude and photoperiod could indeed affect the evolution of circadian clock genes.

The selection mechanisms we could investigate at the genomic level are clearly unlikely to affect all clock genes, and even less so to act with the same intensity on most of them. The emerging picture seems one in which local selective pressures had an effect on one or a few components of the human biological clock. However, these effects are likely minor, as shown by the absence of significant differences in the global patterns of variation for clock and clock-related genes on the

one hand, and the set of presumably neutral markers we analyzed on the other. Larger samples of populations will certainly help telling apart the effects of positive selection from those of shared demographic history. However, we feel that an even more crucial leap forward will be represented by the development of new methods allowing one to jointly consider several genes and their interactions, and compare them with a large set of variables in the environment.

DATA ARCHIVING

Data available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.pr078>.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by a grant of the Italian Ministry for Research and Universities (MIUR) PRIN 2010–2011, by the European Research Council ERC-2011-AdG_295733 grant (LanGeLin) and by a University of Ferrara Research Grant. We are indebted to Paul Verdu for help in the analyses, and Rudi Costa for critical reading of the manuscript.

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO et al. (2015). A global reference for human genetic variation. *Nature* **526**: 68–74.

Alexander DH, Novembre J, Lange K (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**: 1655–1664.

Allebrandt KV, Teder-Laving M, Akylou M, Pichler I, Müller-Myhsok B, Pramstaller P et al. (2010). CLOCK gene variants associate with sleep duration in two independent populations. *Biol Psychiatry* **67**: 1040–1047.

Arbiza L, Zhong E, Keinan A (2012). NRE: a tool for exploring neutral loci in the human genome. *BMC Bioinformatics* **14**: 301.

Archer SN, Carpen JD, Gibson M, Lim GH, Johnston JD, Skene DJ et al. (2010). Polymorphism in the PER3 promoter associates with diurnal preference and delayed sleep phase disorder. *Sleep* **33**: 695–701.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA et al. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* **74**: 1111–1120.

Bertolucci C, Cavallari N, Colognesi I, Aguzzi J, Chen Z, Caruso P et al. (2008). Evidence for an overlapping role of CLOCK and NPAS2 transcription factors in liver circadian oscillators. *Mol Cell Biol* **28**: 3070–3075.

Cavalli-Sforza LL (1966). Population structure and human evolution. *Proc R Soc* **164**: 362–379.

Ciarleglio CM, Ryckman KK, Servick SV, Hida A, Robbins S, Wells N et al. (2008). Genetic differences in human circadian clock genes among worldwide populations. *J Biol Rhythms* **23**: 330–340.

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D et al. (2009). The role of geography in human adaptation. *PLoS Genet* **5**: e1000500.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185**: 1411–1423.

Cruciani F, Trombetta B, Labuda D, Modiano D, Torroni A, Costa R et al. (2008). Genetic diversity patterns at the human clock gene Period 2 are suggestive of population-specific positive selection. *Eur J Hum Genet* **16**: 1526–1534.

Excoffier L, Hofer T, Foll M (2009). Detecting loci under selection in a hierarchically structured population. *Heredity (Edinb)* **103**: 285–298.

Excoffier L, Lischer HEL (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Res* **10**: 564–567.

Foll M, Gaggiotti O (2008). A genome scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**: 977–993.

Forni D, Pozzoli U, Cagliari R, Tresoldi C, Menozzi G, Riva S et al. (2014). Genetic adaptation of the human circadian clock to day-length latitudinal variations and relevance for affective disorders. *Genome Biol* **15**: 499.

Gunther T, Coop G (2013). Robust identification of local adaptation from allele frequencies. *Genetics* **195**: 205–220.

Hamblin MT, Di Rienzo A (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* **66**: 1669–1679.

Harris EE, Meyer D (2006). The molecular signature of selection underlying human adaptations. *Am J Phys Anthropol* **49**: 89–130.

Hatori M, Panda S (2010). The emerging roles of melanopsin in behavioral adaptation to light. *Trends Mol Med* **16**: 435–446.

Hida A, Kitamura S, Katayose Y, Kato M, Ono H, Kadotani H et al. (2014). Screening of clock gene polymorphisms demonstrates association of a PER3 polymorphism with morningness-eveningness preference and circadian rhythm sleep disorder. *Sci Rep* **4**: 6309.

Hu Y, Shmygelska A, Tran D, Eriksson N, Tung JY, Hinds DA (2016). GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. *Nat Commun* **7**: 10448.

Huber W, Carey WG, Gentleman R, Anders S, Carlson M, Carvalho BS et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**: 115–121.

Hut RA, Beersma DG (2011). Evolution of time-keeping mechanisms: early emergence and adaptation to photoperiod. *Philos Trans R Soc Lond B Biol Sci* **366**: 2141–2154.

Jagannath A, Hughes S, Abdelgany A, Potheary CA, Di Pretoro S, Pires SS et al. (2015). Isoforms of melanopsin mediate different behavioral responses to light. *Curr Biol* **25**: 2430–2434.

Jakobsson M, Rosenberg NA (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**: 1801–1806.

Jaramillo-Correa JP, Rodríguez-Quilón I, Grivet D, Lepoittevin C, Sebastiani F, Heuertz M et al. (2015). Molecular proxies for climate maladaptation in a long-lived tree (*Pinus pinaster* Aiton, Pinaceae). *Genetics* **199**: 793–807.

Jeffreys H (1961). *Theory of Probability*. Clarendon Press.

Kyriacou CP, Peixoto AA, Sandrelli F, Costa R, Tauber E (2008). Clines in clock genes: fine-tuning circadian rhythms to the environment. *Trends Genet* **24**: 124–132.

Lane JM, Vlasac I, Anderson SG, Kyle SD, Dixon WG, Bechtold DA et al. (2016). Genome-wide association analysis identifies novel loci for chronotype in 100,420 individuals from the UK Biobank. *Nat Commun* **7**: 10889.

Lee H, Chen R, Lee Y, Yoo S, Lee C (2009). Essential roles of CK1delta and CKIepsilon in the mammalian circadian clock. *Proc Natl Acad Sci USA* **106**: 21359–21364.

Lee HJ, Kim L, Kang SG, Yoon HK, Choi JE, Park YM et al. (2011). PER2 variation is associated with diurnal preference in a Korean young population. *Behav Genet* **41**: 273–277.

Li J, Li H, Jakobsson M, Li S, Sjödin P, Lascoux M (2012). Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Mol Ecol* **21**: 28–44.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**: 1100–1104.

Lucas RJ (2013). Mammalian inner retinal photoreception. *Curr Biol* **23**: R125–R133.

Mohawk JA, Green CB, Takahashi JS (2012). Central and Peripheral Circadian Clocks in Mammals. *Annu Rev Neurosci* **35**: 445–462.

Novembre J, Di Rienzo A (2009). Spatial patterns of variation due to natural selection in humans. *Nat Rev Genet* **10**: 745–755.

O'Malley KG, Banks MA (2008). A latitudinal cline in the Chinook salmon (*Oncorhynchus tshawytscha*) Clock gene: evidence for selection on PolyQ length variants. *Proc Biol Sci* **275**: 2813–2821.

Parra EJ (2007). Human pigmentation variation: evolution, genetic basis, and implications for public health. *Am J Phys Anthropol (Suppl)* **45**: 85–105. Review.

Partch CL, Green CB, Takahashi JS (2014). Molecular architecture of the mammalian circadian clock. *Trends Cell Biol* **24**: 90–99.

Partonen T, Treutlein J, Alpmann A, Frank J, Johansson C, Depner M et al. (2007). Three circadian clock genes *Per2*, *Arntl*, and *Npas2* contribute to winter depression. *Ann Med* **39**: 229–238.

Patin E, Laval G, Barreiro LB, Salas A, Semino O, Santachiara-Benerecetti S et al. (2009). Inferring the demographic history of African Farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet* **5**: e1000448.

Pendergast JS, Friday RC, Yamazaki S (2010). Distinct functions of Period2 and Period3 in the mouse circadian system revealed by in vitro analysis. *PLoS One* **5**: e8552.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D et al. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* **81**: 559–575.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA* **102**: 15942–15947.

Reick M, Garcia JA, Dudley C, McKnight SL (2001). NPAS2: an analog of clock operative in the mammalian forebrain. *Science* **293**: 506–509.

Roeklein KA, Wong PM, Franzen PL, Hasler BP, Wood-Vasey WM, Nimgaonkar VL et al. (2012). Melanopsin gene variations interact with season to predict sleep onset and chronotype. *Chronobiol Int* **29**: 1036–1047.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA et al. (2002). Genetic structure of human populations. *Science* **298**: 2381–2385.

Sato TK, Panda S, Miraglia LJ, Reyes TM, Rudic RD, McNamara P et al. (2004). A functional genomics strategy reveals Rora as a component of the mammalian circadian clock. *Neuron* **43**: 527–537.

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G et al. (2001). Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* **293**: 455–462.

Xu Y, Padiath QS, Shapiro RE, Jones CR, Wu SC, Saigoh N et al. (2005). Functional consequences of a CK1delta mutation causing familial advanced sleep phase syndrome. *Nature* **434**: 640–644.

Zhou M, Kim JK, Eng GW, Forger DB, Virshup DM (2015). A Period2 Phosphoswitch regulates and temperature compensates circadian period. *Mol Cell* **60**: 77–88.

Zhu Y, Stevens RG, Leaderer D, Hoffman A, Holford T, Zhang Y et al. (2008). Non-synonymous polymorphisms in the circadian gene NPAS2 and breast cancer risk. *Breast Cancer Res Treat* **107**: 421–425.