

# Targeted resequencing of two genes, *RAGE* and *POLL*, confirms findings from a genome-wide scan for adaptive evolution and provides evidence for positive selection in additional populations

Joanna L. Kelley\*, Kayley Turkheimer, Margo Haney and Willie J. Swanson

Department of Genome Sciences, University of Washington, Seattle, WA, USA

Received August 23, 2008; Revised and Accepted November 19, 2008

---

The availability of multiple genome-wide human polymorphism datasets has led to an increase in efforts to scan the genome for signals of positive selection. As a result, the number of loci in the human genome predicted to be adaptively evolving increases monthly. Yet, these numerous genome-wide scans have identified minimally overlapping sets of candidate loci, potentially due to biases in genotype versus sequence data or power of statistical tests to detect selection in different time frames. Because of these issues, a critical step is to confirm the evidence for positive selection through direct sequencing. In this study, we describe the resequencing and analysis of two loci, *RAGE* and *POLL*, that were identified by a recent genome-wide scan of the Perlegen data to be under selection in the Han Chinese population. By resequencing these loci in additional populations, we have found that the evolutionary history of these regions is more complex than observed in the initial genome-wide scan and that the sweep patterns are shared across several populations. The resequencing data provide evidence for selection on *RAGE* in the non-African populations and on *POLL* in the Asian and Sub-Saharan African populations. In addition to confirming the signatures of selection from the genome-wide scan, direct resequencing reveals more extensive patterns of selection than the genotype data.

---

## INTRODUCTION

Genome-wide scans have identified numerous candidate loci predicted to have been targets of positive selection (1–15). These loci are now the focus of additional investigation. While some researchers have begun to address reproducibility across the different genome-wide scans, few have followed up the predictions from genotype data with targeted resequencing of the genes. This is an important step in understanding the selective pressure acting upon these loci because the majority of statistical tests were developed to be applied to sequence data rather than genotype data. In addition, it is unknown whether more complex patterns will emerge from the analysis of sequence data across diverse populations. Direct sequencing of adaptively evolving loci is a first step towards the verification of functional differences in a laboratory setting.

The abundance of genotype data available has led to an increase in the ability to detect selection in the human genome. An outcome of the Perlegen Sciences and the

International HapMap Project efforts was the availability of extensive single-nucleotide polymorphism (SNP) genotype data for population genetic analysis (5,10,16). Genome-wide SNP genotype data provide a way to compare population genetic statistics across the entire genome. Demographic events will distort the population genetic signatures genome wide; by analyzing the entire genomic dataset, it is possible to create empirical distributions of population genetic measures with the hope that the outliers represent true positives. The methods used to scan the genome for evidence of positive selection vary from categorizing regions of low diversity (2,11) to identifying extended haplotypes that diverge from the expected neutral haplotype patterns (3,4,13,14). When all genome-wide scans are compared, it appears that there is minimal overlap between the scans. Each scan identifies tens to hundreds of candidate loci; however, only a fraction of those identified appear in any other scan (17,18).

The genotype data used in genome-wide scans have inherent biases. The methods used to collect an individual's genotype, the

---

\*To whom correspondence should be addressed. Tel: +1 7738349836; Fax: +1 7738340505; Email: jlkelly@uchicago.edu

population panel used for the initial ascertainment that led to the preferential inclusion of high-frequency SNPs, and the addition of SNPs to increase the map density all create biases that must be taken into account when using population genetic statistics. These biases may alter the statistical distribution(s) in a way similar to the action of selection. For example, by ascertaining for common alleles, population genetic statistics that rely on the allele frequency will be erroneously shifted toward results that reflect an excess of common alleles, indicative of balancing selection. An approach researchers use to address the described biases is to create an empirical distribution of the chosen test statistic. By selecting the loci in the tail of the empirical distribution, it is possible to identify candidate selection loci without a priori assumptions about the nature of the selective pressure. However, it remains unknown how SNP genotyping biases affect the empirical distribution.

The first step in connecting evidence for positive selection from genome-wide scans with specific gene function is to confirm the evidence of selection by direct sequencing. Because we do not have a complete picture of which genes are true examples of selected loci, it is difficult to estimate the false-positive rate for any genome-wide scan. In an attempt to address this problem, it has been shown that a modified Tajima's D calculated using genotype data correlates with Tajima's D calculated directly from sequence data (8,11). Although researchers do their best to minimize the false-positive rate, the outliers, or loci in the tail of the distribution, contain both true and false positives. Direct sequencing is one potential way to separate the true positives from false positives. Ultimately, biological studies provide a way to assess the functional significance of the adaptive changes. Even so, in regards to biological studies, the past selective pressure may no longer be present, thereby making it impossible to replicate, understand or estimate the functional difference between the selected and non-selected alleles. The allele frequency spectrum and haplotype structure will help confirm whether or not the loci have been subject to recent positive selection. If the sequence data confirm the evidence for positive selection, additional sequence data from multiple populations will allow researchers to estimate whether or not the loci were adaptively evolving in other populations. It will also provide complete data to estimate the timing of selective events.

We previously identified 385 genes that have polymorphism patterns consistent with positive selection (11). This genome-wide scan analyzed the three Perlegen populations: Han Chinese, African-Americans and European-Americans, for evidence of a recent complete selective sweep. The scan utilized an outlier approach to identify loci with decreased nucleotide variation and an excess of rare alleles, which was measured by a modified Tajima's D statistic. When an adaptive mutation appears in a population, it is brought to high frequency due to the increased fitness of those individuals who carry the allele. As the allele rises in frequency in the population, the surrounding neutral region is also swept to high frequency, reducing neutral variation in the region. After completion of a selective sweep, new mutations accumulate in the swept region leading to an excess of rare alleles, which can be assayed for by genome-wide scans. However, population demographic events may also lead to similar deviations in the allele frequency spectrum. Genome-wide scans based on genotype

data cannot rule out population demographics as the underlying mechanism for observed polymorphism patterns. The effects of demographic events can be ruled out by combining the empirical distributions with simulations and comparisons to non-gene regions. Direct resequencing provides additional data to confirm the signature of selection and to begin follow-up studies to genome-wide scans. In this study, we present resequencing data from two genes, renal tumor antigen (*RAGE*) and DNA polymerase lambda (*POLL*) identified in a genome-wide scan (11).

## RESULTS

Both the *RAGE* and *POLL* loci show evidence for positive selection only in the Han Chinese (CA) Perlegen population, thus allowing us to verify two of the outliers from the empirical distribution of Tajima's D for this population (11). Also, within the coding sequence of each locus is a nonsynonymous polymorphism, which could have been the target of the respective selective sweeps. The presence of a nonsynonymous SNP within the selected region gives us a starting point for exploring the biological basis for selection (Table 1). The genes were not chosen based on nonsynonymous allele frequency differences between populations. The nonsynonymous-derived allele in *RAGE* is at high frequency; the presence of high-frequency-derived alleles is consistent with positive selection (19,20). However, the nonsynonymous allele in *POLL* has a low-frequency-derived allele, which is consistent with relaxed purifying selection. Furthermore, each locus was the only one in the surrounding chromosomal region with evidence for positive selection. By choosing loci that were not surrounded by other genes with evidence for positive selection, it is more likely that positive selection had acted on or near the locus and helps rule out the possibility that the unusual signal was the result of genetic hitchhiking during a selective sweep from a nearby gene.

The goal of this study was to verify the signatures of selection identified in these two loci in a genome-wide scan and to expand the analysis to a worldwide panel of individuals. We chose nine populations for resequencing (see Materials and Methods). While the individuals chosen for resequencing differ from those genotyped by Perlegen, the individuals were chosen from similar populations, with an emphasis on populations that were not admixed. Additional populations representing a broader range of populations across the globe were also included for resequencing.

From the genome-wide polymorphism patterns, *RAGE* is predicted to have evolved adaptively only in the Han Chinese population. A total of 12 SNPs were genotyped by Perlegen, and in the Han Chinese population four of the SNPs were absent and the other eight have a minor allele frequency of <5%. We resequenced 3680 bp of the *RAGE* locus, targeting exonic regions. Eleven SNPs were identified and characterized in our sequencing efforts, three of which were also genotyped by Perlegen (Supplementary Material, Table S1). For the analyses, the nine populations were combined into five major groups based on clustering methods performed by Rosenberg *et al.* (21) using 377 microsatellites. In the combined Asian population, Tajima's D and Fay and Wu's H are significantly negative, indicating departure from

**Table 1.** Perlegen-derived allele frequencies for nonsynonymous SNPs in *RAGE* and *POLL*

Gene	SNP	AA	EA	CA	Amino acid change
RAGE	rs2236493	0.543	0.750	0.958	R→Q
POLL	rs3730477	0.022	0.146	0.021	R→W

AA, African American; EA, European American; CA, Han Chinese.

equilibrium neutral expectation, possibly due to positive selection (Table 2). Tajima's D compares nucleotide polymorphism and nucleotide diversity to identify regions that have higher or lower nucleotide diversity than neutral expectations (22). Whereas, Fay and Wu's H analyzes the relative number of derived, non-ancestral alleles and the allele frequency of such alleles to identify hitchhiked regions (19). There are two SNPs whose derived alleles are at high frequency in the European, South American and Asian populations (Fig. 1A and Table 2, Fay and Wu's H). While the polymorphism patterns are suggestive of a selective sweep, we cannot definitively conclude that selection has been acting on *RAGE* in the Asian population. Selection and/or demography can lead to the presence of derived allele(s) at high frequency in the non-African populations. However, the results from our resequencing efforts confirm the Perlegen genome-wide scan findings (11) for positive selection potentially acting in the Han Chinese population and extend the signal to the European and South American populations, but not African populations.

Similar to *RAGE*, *POLL* was identified as a candidate gene under selection in the genome-wide scan and had evidence of a selective sweep only in the Han Chinese population. On the basis of allele frequencies from the 10 SNPs genotyped by Perlegen in the *POLL* locus, the African-American and Han Chinese populations have low levels of variation while the European-American population SNPs are at an intermediate frequency. We characterized nine SNPs in the *POLL* locus, five of which were genotyped by Perlegen (Fig. 1B). The SNPs occur on two major haplotypes, here called H1 and H2. H1 has three derived alleles, whereas H2 has one derived allele and a derived, 5-bp deletion (Table 3). The two haplotypes are each more closely related to the chimpanzee (ancestral) haplotype than to each other. Additionally, the two haplotypes are present in all five populations, though H2 is often at low frequency. One of the derived H1 alleles is located in an exon and encodes a synonymous substitution. While this SNP is unlikely to be functional, a recent study suggested that synonymous changes could affect protein structure (23,24). The H1 haplotype is found at high frequency in the Asian, Sub-Saharan African and South American populations. The near fixation of a derived haplotype (H1) in Sub-Saharan Africa and its high frequency in Asia is consistent with adaptive evolution. Tajima's D and Fay and Wu's H are significantly negative in the Asian and Sub-Saharan African populations, indicating that the region has been subject to positive selection in those populations (Table 1) and that the H1 haplotype contains the sites subjected to positive selection. Owing to genotyping of African-Americans by Perlegen, signals of selection in the African populations may be hidden by population admixture.

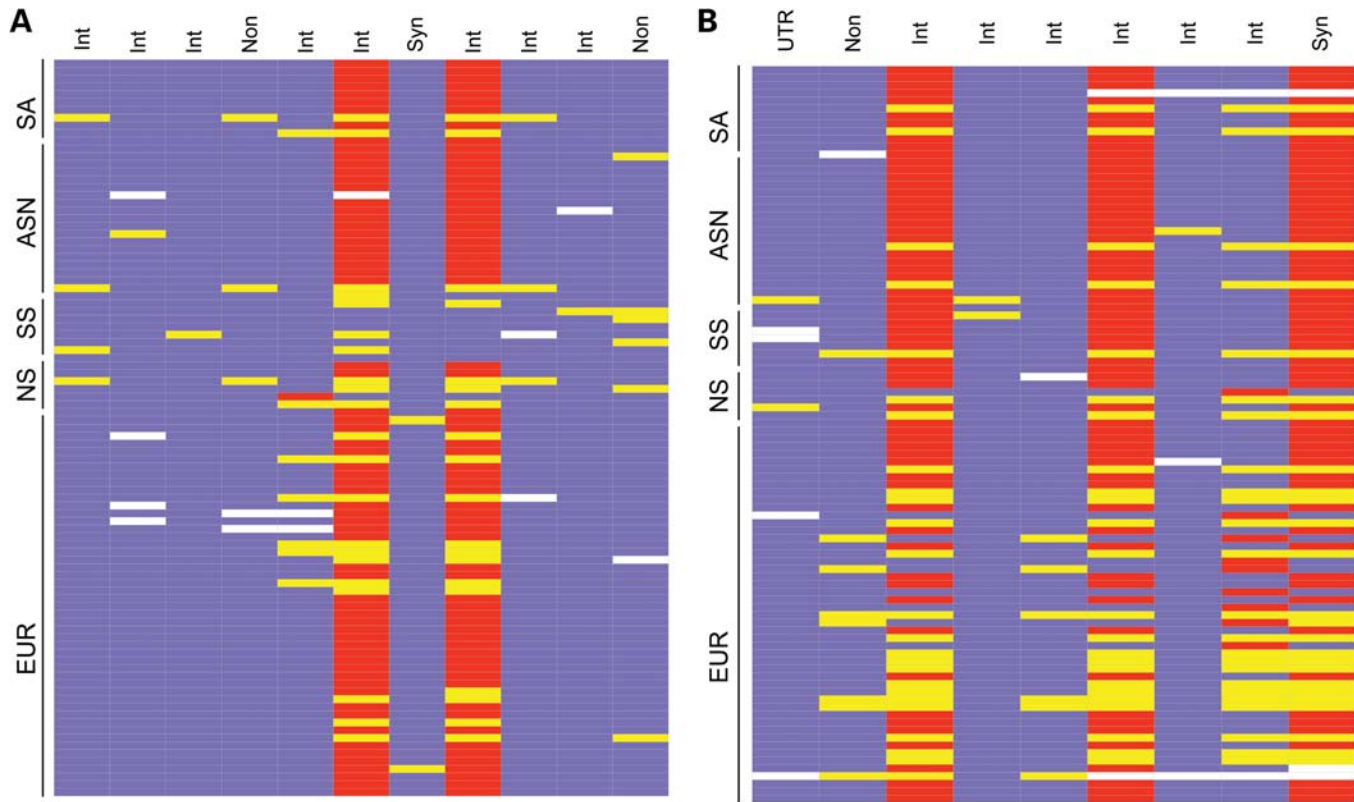
## DISCUSSION

In addition to verifying the signals of selection identified in a genome-wide scan, the resequencing of *RAGE* and *POLL* supports the hypothesis of population-specific adaptation. Polymorphism patterns differ between populations and between the genes. These genes were not identified by other scans for positive selection, below are hypotheses of why other genome-wide scans failed to find the signature of selection at these loci. Several of the scans were developed to identify evidence of a selective sweep on a different timescale; for example, test-statistics based on haplotype homozygosity are limited to identifying incomplete selective sweeps (13,14). The scan executed by Carlson *et al.* (8) was developed to identify genes with reduced Tajima's D calculated from SNP data; however, the method required very large genomic regions of at least 300 kb to share the signal of selection. Both the Williamson and Kimura scans are aimed at detecting complete sweeps, so it is unclear why they did not identify *RAGE* and *POLL*.  $d_N/d_S$  methods rely on fixed amino acid substitutions and assume that selection has been acting on a longer timescale (25). Repeating nonsynonymous changes at a site leads to a higher rate of nonsynonymous substitutions ( $d_N$ ) than synonymous substitutions ( $d_S$ ). This results in a  $d_N/d_S$  ratio  $>1$  at that site, which is consistent with positive selection.

On the basis of our results, when compared with genome-wide analysis of genotypic data, targeted resequencing reveals more complex evolutionary histories at loci predicted to be under positive selection than genotype data. While this study has verified the signals of selection, linkage disequilibrium (LD) in the region precludes identifying the mutational target of selection. Both *RAGE* and *POLL* are located in blocks of LD within the African-American and European-American populations. However, the minor allele frequency is so low in the Han Chinese population in both genes, there is not enough information to measure LD. Linkage in the region, even with resequencing data, it makes difficult to identify the mutation that was the target of selection. Future collaborations may further elucidate how these mutations alter the function or expression of *RAGE* and *POLL* and shed light on the causes behind the recent selection identified by the genome-wide scan and confirmed in this study.

## MATERIALS AND METHODS

We targeted sequencing to exonic regions of *RAGE* and *POLL* in 95 and 96 individuals, respectively, from nine populations. We focused on coding polymorphisms, although selection could also act on non-coding regions, for example, affecting gene expression patterns (26). DNA was from Coriell Cell Repositories Human Variation Collection, numbers in parentheses correspond to the number of individuals sequenced from the each panel: Northern European HD01 (NA17002-17010) (9), Russian HD23 (10), Africans North of the Sahara HD11 (7), Africans South of the Sahara HD12 (9), Middle East HD05 (Version 1) (10), Japanese HD07 (10), Aboriginal Tribe from Taiwan HD24 (10), South America HD17 (10) and Caucasian HD50CAU (NA17231-17250) (20). Sequences are deposited in dbSNP.



**Figure 1.** Visual genotype of *RAGE* (A) and *POLL* (B). Each row represents an individual, rows are grouped by population. Each column represents a polymorphic site. The nine sequenced populations were clustered into five geographic locations, indicated on the left as: EUR = European, NS = North Saharan Africans, SS = Sub-Saharan Africans, ASN = Asians and SA = South Americans. Red indicates homozygous for the derived allele, Yellow indicates heterozygous and blue indicates homozygous for the ancestral allele, determined by comparison to the chimpanzee genome. White indicates missing data. The location of each SNP (Int = intronic, Syn = synonymous, Non = nonsynonymous or UTR = untranslated region) is noted above each column.

**Table 2.** Population genetic statistics for *RAGE* and *POLL* resequence data

Gene	Population	bp, analyzed	Sample size	<i>S</i>	<i>k</i>	Tajima's D	Fay and Wu's H
<i>RAGE</i>		<b>3680</b>	<b>190</b>	<b>11</b>			
	European		98	5	0.572	-0.875	-2.746*
	South American		20	6	0.787	-1.076	-2.655*
	Asian		40	7	0.345	-2.189*	-3.508*
	N. Saharan		14	7	1.923	-0.466	0.000
	S. Saharan		18	6	1.050	-1.208	0.850
<i>POLL</i>		<b>4001</b>	<b>190</b>	<b>9</b>			
	European		100	6	1.985	1.602	-0.945
	South American		18	4	0.837	-0.841	-4.209*
	Asian		40	5	0.450	-1.593*	-5.076*
	N. Saharan		14	5	1.901	0.724	-1.062
	S. Saharan		18	6	0.765	-1.849*	-4.527*

Sample size is number of chromosomes. Values in bold and with asterisks indicate  $P < 0.05$ . *S*, number of segregating sites; *k*, average number of nucleotide differences.

## RAGE

We sequenced 3680 bp of the *RAGE* locus, which spans 76 kb. Primers were designed to amplify the exons, using PRIMER3 v0.2 (27), and the known human sequence, reference sequence NM\_014226 (UCSC Genome Browser March 2006 Assembly). The primers and PCR conditions are available upon request. The PCR products were analyzed on an ABI 3100

automated sequencer after  $5 \times$  dilution, cycle-sequencing using BigDye, v. 3.1, and ethanol precipitation.

## POLL

Of the *POLL* locus, 4001 bp were sequenced. The *POLL* locus spans 9.34 kb. Primers were designed to amplify the exons,

**Table 3.** POLL haplotypes

Chimp	CAACA	A	A	C	T	C	G	C	T	C
Haplotype1	CAACA	A	A	T	T	C	A	C	T	T
Haplotype2		A	A	C	T	C	G	C	C	C

The grey boxes indicate derived allele.

using PRIMER3 v0.2 (27), and the annotated human sequence, reference sequence NM\_013274 (UCSC Genome Browser March 2006 Assembly). Primers and PCR conditions are available upon request.

Sequence data for each locus were independently base-called, assembled and scanned for polymorphisms using Phred, Phrap and polyPhred (28–30). Polymorphisms were visually confirmed using Consed (31). The assembled and inspected sequences were exported to infer haplotype using PHASE (32). Population genetic parameter estimations and tests of neutrality were executed using Arlequin (33) and DnaSP, v4.0 (34).

Population genetic tests of neutrality: to determine whether selection was acting on either locus in one or more population we used the statistical tests Tajima's D (22) and Fay and Wu's H (19). Both tests are sensitive to deviations from neutral expectations, as defined by the neutral theory of evolution (35). Fay and Wu's H was calculated by DnaSP, v4.0, disregarding missing sites; the data were also analyzed disregarding individuals with missing data, the results were consistent. Tajima's D was calculated using Arlequin for sites with less than 5% missing data.

## SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

## ACKNOWLEDGEMENTS

We thank Geoff Findlay and Carole Kelley for comments on the manuscript.

*Conflict of Interest statement.* None declared.

## FUNDING

This work was supported by a Sigma Xi grant in aid of research [to J.L.K.]; the National Science Foundation [DDIG DEB-0709660 to J.L.K. and DEB-0716761 to W.J.S.]; and National Institutes of Health [HD057974, HD054631, HD042563 to W.J.S.].

## REFERENCES

- Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsepas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R. *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.
- Williamson, S.H., Hubisz, M.J., Clark, A.G., Payseur, B.A., Bustamante, C.D. and Nielsen, R. (2007) Localizing recent adaptive evolution in the human genome. *PLoS Genet.*, **3**, e90.
- Kimura, R., Fujimoto, A., Tokunaga, K. and Ohashi, J. (2007) A practical genome scan for population-specific strong selective sweeps that have reached fixation. *PLoS ONE*, **2**, e286.
- Tang, K., Thornton, K.R. and Stoneking, M. (2007) A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.*, **5**, e171.
- International\_HapMap\_Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
- Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A. and Kruglyak, L. (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.*, **2**, e286.
- Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D. *et al.* (2005) Natural selection on protein-coding genes in the human genome. *Nature*, **437**, 1153–1157.
- Carlson, C.S., Thomas, D.J., Eberle, M.A., Swanson, J.E., Livingston, R.J., Rieder, M.J. and Nickerson, D.A. (2005) Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.*, **15**, 1553–1565.
- Clark, A.G., Glanowski, S., Nielsen, R., Thomas, P., Kejariwal, A., Todd, M.J., Tanenbaum, D.M., Civello, D., Lu, F., Murphy, B. *et al.* (2003) Positive selection in the human genome inferred from human-chimp-mouse orthologous gene alignments. *Cold Spring Harb. Symp. Quant. Biol.*, **68**, 471–477.
- International\_HapMap\_Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
- Kelley, J.L., Madeoy, J., Calhoun, J.C., Swanson, W. and Akey, J.M. (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.*, **16**, 980–989.
- Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J. *et al.* (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.*, **3**, e170.
- Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.
- Voight, B.F., Kudravalli, S., Wen, X. and Pritchard, J.K. (2006) A map of recent positive selection in the human genome. *PLoS Biol.*, **4**, e72.
- Wang, E.T., Kodama, G., Baldi, P. and Moyzis, R.K. (2006) Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proc. Natl Acad. Sci. USA*, **103**, 135–140.
- Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A. and Cox, D.R. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072–1079.
- Biswas, S. and Akey, J.M. (2006) Genomic insights into positive selection. *Trends Genet.*, **22**, 437–446.
- Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. and Clark, A.G. (2007) Recent and ongoing selection in the human genome. *Nat. Rev. Genet.*, **8**, 857–868.
- Fay, J.C. and Wu, C.I. (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.
- Fay, J.C., Wyckoff, G.J. and Wu, I. (2001) Positive and negative selection on the human genome. *Genetics*, **158**, 1227–1234.
- Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. and Feldman, M.W. (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.
- Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
- Drummond, D.A. and Wilke, C.O. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, **134**, 341–352.
- Powers, E.T. and Balch, W.E. (2008) Costly mistakes: translational infidelity and protein homeostasis. *Cell*, **134**, 204–206.
- Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D. and Lander, E.S. (2006) Positive natural selection in the human lineage. *Science*, **312**, 1614–1620.

26. Gilad, Y., Oshlack, A. and Rifkin, S.A. (2006) Natural selection on gene expression. *Trends Genet.*, **22**, 456–461.
27. Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
28. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
29. Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
30. Nickerson, D.A., Tobe, V.O. and Taylor, S.L. (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.*, **25**, 2745–2751.
31. Gordon, D., Abajian, C. and Green, P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.
32. Stephens, M., Smith, N.J. and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.
33. Excoffier, L., Laval, G. and Schneider, S. (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evol Bioinform Online*, **1**, 47–50.
34. Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X. and Rozas, R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.
35. Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.