BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 16 2010, pages 1999–2003
doi:10.1093/bioinformatics/btq327

Genetics and population analysis Advance Access publication June 16, 2010

A Bayesian approach using covariance of single nucleotide polymorphism data to detect differences in linkage disequilibrium patterns between groups of individuals

Taane G. Clark[1,2,*], Susana G. Campino[2], Elisa Anastasi[2], Sarah Auburn[2], Yik Y. Teo[3], Kerrin Small[3], Kirk A. Rockett[3], Dominic P. Kwiatkowski[2,3] and Christopher C. Holmes[4]

[1]Departments of Epidemiology and Public Health and Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, [2]Wellcome Trust Sanger Institute, Hinxton, Cambridge, [3]Wellcome Trust Centre for Human Genetics, University of Oxford and [4]Department of Statistics, University of Oxford, Oxford, UK

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Quantifying differences in linkage disequilibrium (LD) between sub-groups can highlight genetic regions or sites under selection and/or associated with disease, and may have utility in trans-ethnic mapping studies.

Results: We present a novel pseudo Bayes factor (PBF) approach that assess differences in covariance of genotype frequencies from single nucleotide polymorphism (SNP) data from a genome-wide study. The magnitude of the PBF reflects the strength of evidence for a difference, while accounting for the sample size and number of SNPs, without the requirement for permutation testing to establish statistical significance. Application of the PBF to HapMap and Gambian malaria SNP data reveals regional LD differences, some known to be under selection.

Availability and implementation: The PBF approach has been implemented in the BALD (Bayesian analysis of LD differences) C++ software, and is available from http://homepages.lshtm.ac.uk/tgclark/downloads

Contact: taane.clark@lshtm.ac.uk

Received on March 15, 2010; revised on June 10, 2010; accepted on June 11, 2010

1 INTRODUCTION

Comparing linkage disequilibrium (LD) patterns between groups, such as cases and controls or sub-populations, can provide an insight into genetic regions under selection and/or association with disease. In fact, comparisons between different ethnic groups may be used to calibrate the resolution of association using trans-ethnic mapping (McKenzie *et al.*, 2001), as well as indicate the appropriateness of genotypic imputation using sequence data from a reference (HapMap) panel. A number of approaches have been suggested for quantifying LD differences. *Haplotype sharing* methods (Te Meerman and Van der Meulen, 1997; Thomas *et al.*, 2003; Tzeng *et al.*, 2003) are usually applied in a case–control setting, where it is assumed that pairs of cases would tend to be more closely related than pairs of controls, while case–control pairs would be even more distantly related on average. Such approaches isolate regions of interest using pairwise comparisons of lengths of haplotype sharing (Bourgain *et al.*, 2001; Wang *et al.*, 2006), they can be generalized to compare different sub-populations, and use U-statistics and permutations to determine statistical significance. Other methods compare the matrices of pairwise LD metrics (e.g. r^2 and D') of different sub-populations. In particular, they calculate differences in the eigenvalues (and eigenvectors) (Krzanowski, 1993; Teo *et al.*, 2009; Zaykin *et al.*, 2006), and perform permutations to determine statistical significance thresholds. Here, we present an alternative approach that does not require permutations and is based on analysis of covariance of genotype frequencies. We propose a Bayes factor statistic that quantifies the degree to which covariance matrices from sub-populations are different. We compare our method with an eigenvalue difference (EVD) approach, and apply it to data from the HapMap (The International HapMap Consortium, 2007) and a subset of controls from a genome-wide association study (GWAS) of malaria in a Gambian population (Jallow *et al.*, 2009).

2 METHODS

Assume X and Y are allele counts for p SNPs, which are individually binomial distributed. When the number of individuals (n) is large (as in a GWAS), the multivariate central limit theorem implies the Gaussian distributions $\sqrt{n_x}(\overline{X}-\mu_x)\rightarrow N(0,\Sigma_X)$ and $\sqrt{n_y}(\overline{Y}-\mu_y)\rightarrow N(0,\Sigma_Y)$, where $\mu_.$ and $\Sigma_.$ are mean vectors (length p) and covariance matrices (dimension $p\times p$) respectively, and $\overline{X}$ and $\overline{Y}$ are the sample means. Suppose we wish to test the hypothesis that the covariances are the different ($H_1:\Sigma_X\neq\Sigma_Y$) versus that they are the same ($H_0:\Sigma_X=\Sigma_Y=\Sigma_{X,Y}$). To do this in a Bayesian framework, we simply specify two models for the data, one which assumes that the samples have a common covariance and the other where two covariances are needed. We then calculate the posterior probability under the two models, assuming equal prior probabilities of the models,which is directly related to the Bayes factor, the ratio of the marginal likelihoods. In other words, we compare the model in $H_0:Pr(X,Y)$ versus that in $H_1:Pr(X)Pr(Y)$. More explicitly the Bayes factor is given by

$$\frac{Pr(X,Y|\mu_X,\mu_Y)}{Pr(X|\mu_X)Pr(Y|\mu_Y)} \quad (1)$$

with

$$Pr(X,Y|\mu_X,\mu_Y)=\int Pr(X,Y|\Sigma_{X,Y},\mu_X,\mu_Y)\pi(\Sigma_{X,Y})d\Sigma$$

*To whom correspondence should be addressed.

where $Pr(X,Y|\Sigma_{X,Y},\mu_X,\mu_Y)$ is the data likelihood, $\pi(\cdot)$ denotes a prior distribution on the covariance matrix and

$$Pr(X|\mu_X)Pr(Y|\mu_Y) = \int Pr(X|\Sigma_X,\mu_X)\pi(\Sigma_X)d\Sigma_X \times \int Pr(Y|\Sigma_Y,\mu_Y)\pi(\Sigma_Y)d\Sigma_Y.$$

Kass and Raftery (1995) suggest a framework for interpreting the magnitude of a Bayes factor and, therefore, the level of evidence against H_0. For Gaussian distributed data (X,Y) it is convenient to adopt the conjugate prior for the covariance Σ, which is the inverse-Wishart distribution (Gelman *et al*., 2003), specifically,

$$\pi(\Sigma|d,D) = I_w[d,D]|\Sigma^{-1}|^{-(d+p+1)/2}\exp\{-\frac{1}{2}\mathrm{tr}(D\Sigma^{-1})\} \quad (2)$$

where d is a shape scalar, D the $p \times p$ *inverse* scale matrix, *tr* refers to the trace of a matrix and $I_w[d,D]$ the normalizing constant,

$$I_w[d,D] = \frac{|D|^{d/2}}{2^{dp/2}\Gamma_p(d/2)}.$$

The Wishart distribution is the sampling distribution of the matrix of sums of squares and products, and is the multivariate analogue to the χ^2 (Gelman *et al*., 2003). We say that Σ follows an inverse Wishart distribution, if Σ^{-1} follows a Wishart distribution. An inverse-Wishart distribution is a flexible representation (Gelman *et al*., 2003), and it is possible to modify the off-diagonal elements of D to incorporate knowledge of regions (single nucleotide polymorphisms, SNPs) with high LD. Assume Z is the grouped data of X and Y $\{X,Y\}$, with mean μ_Z and covariance Σ_Z. In this case, (1) can be calculated in closed-form, and where $Z \sim N(\mu_Z,\Sigma_Z)$ with prior $\Sigma_Z \sim$ inverse-Wishart(d,D) we have,

$$Pr(Z|\mu_Z) = \frac{1}{(2\pi)^{np/2}}\frac{I_w[d,D]}{I_w[d+n,D+U]} \quad (3)$$

where

$$U = \sum_{i=1}^{n}(z^{(i)}-\mu_Z^{(i)})(z^{(i)}-\mu_Z^{(i)})'$$

and the normalizing constant $I_w[a,A]$ has the same form as above. Our algorithm may be summarized in the following steps:

1. Centre the genotype data by subtracting the sample means for each SNP ($\hat{\mu}_\cdot$): $X \leftarrow X-\hat{\mu}_X$ and $Y \leftarrow Y-\hat{\mu}_Y$.
2. Group data $Z=\{X,Y\}$ and calculate the logarithm of (3).
3. Calculate the logarithm of (3) for X and Y separately, where n is replaced by n_x and n_y (the number of observations or individuals for X and Y), respectively.
4. The log pseudo Bayes factor (PBF) is the difference of Step 2 minus the sum of individual factors in Step 3.

We describe PBF as a pseudo-metric as we rely on various assumptions, such as the equality of the priors and the asymptotic properties for allele counts. More, negative PBF values [greater denominator in (1)] provide greater evidence of a difference in the covariance of genotypes in the sub-populations. Here, we specify the prior parameters in the Wishart$(\Sigma|d,D])$ in a default way, namely $d=p$ (the number of SNPs) and $D=I$ (the identity matrix). Although, our PBF implementation is robust to covariance matrices that are improper or with negative eigenvalues, we suggest that SNPs with high levels of missing genotypes (e.g. $>10\%$) and very low minor allele frequency should be removed from the analysis, in keeping with other genome-wide analysis approaches (Jallow *et al*., 2009). In addition, it is also advisable to remove any SNPs that deviate significantly from Hardy–Weinberg equilibrium (e.g. HWE $P<10^{-7}$), as it may be due to poor genotype calling (Jallow *et al*., 2009). It makes sense to restrict the test to a window of markers and unless otherwise stated, we consider windows of 50 SNPs in our analyses. We compare the PBF to a simple haplotype sharing method (HAPS) that considers in each population, the average number of common alleles across the window of SNPs between all pairwise haplotypes (Tzeng *et al*., 2003), and then calculates the absolute difference between populations. We also compare our approach to a metric based on comparing EVDs [similar to Krzanowski (1993) and Teo *et al*. (2009)] between sub–population matrices of pairwise directional r^2 (Teo *et al*., 2008) values. Specifically, if we assume that these matrices calculated for p SNPs are τ_1 and τ_2 in sub-populations 1 and 2, respectively, and non–improper, then symmetric matrix τ_i can be decomposed into $\Gamma_i\Delta_i\Gamma_i^T$ $(i=1,2)$, where Γ_i is orthogonal and Δ_i diagonal. The EVD is the sum of the absolute values of $\Delta_1-\Delta_2$.

3 SIMULATION STUDY

The coalescent simulation package MS (Hudson, 2002) was used to generate haplotypic data for a 250 kb region assuming an effective population size of 10 000 and a constant mutation rate (μ) of 2.0×10^{-8} per generation per base pair. We considered two levels of LD reflecting different uniform recombination rates (crossover rate r per generation per site): (i) high LD ($r=0.5\times10^{-8}$) and (ii) low ($r=2.0\times10^{-8}$). We generated genetic data for two sub-populations, assuming their haplotypes were from: (i) the same population; (ii) two different populations; and (iii) where one sub-population is a random sample of haplotypes from the two different populations, and the other sub-population consists of a 75 : 25 mixture of haplotypes from those populations. Genotypes were constructed from the haplotypes, and we considered two sample sizes ($n=60250$, identical in each sub-population) and two different SNP densities ($p=50100$). We calculated the EVD, HAPS and PBF values for 2000 replicates of each scenario. The results are presented in Table 1. Decreasing LD leads to lower HAPS, EVD and absolute PBF values. When there are differences in LD, the PBF metric is strictly negative, and its magnitude is greater with

Table 1. The HAPS, EVD and PBF metrics with corresponding 95% credibility regions; the scenarios correspond to identical sub-populations (No), where one sub-population consists of 25% and 75% of the haplotypes of two populations and the other sub-population is a random sample from the two populations (some), and where the haplotype structure in each sub-population is completely different (yes); we assume a constant mutation rate of 2.0×10^{-8} per generation per base pair, and a uniform recombination rate of 2×10^{-8} (low LD) and 0.5×10^{-8} (high LD) per generation per site.

Different	n	p	High LD			Low LD		
			HAPS	EVD	PBF	HAPS	EVD	PBF
No	60	50	0.35	2.89	840.4	0.31	1.99	646.5
			0.02, 1.25	1.82, 4.93	948.5, 1059.0	0.01, 1.03	1.24, 3.38	553.7, 757.4
	60	100	0.57	5.53	2636.7	0.46	3.53	1802.0
			0.02, 2.24	3.47, 9.40	2448.9, 2820.3	0.03, 1.64	2.31, 5.93	1588.8, 2014.1
	250	50	0.17	1.77	2380.4	0.15	1.23	1533.9
			0.01, 0.64	1.19, 2.72	1951.7, 2870.1	0.01, 0.51	0.81, 1.89	1317.7, 1871.8
	250	100	0.31	3.30	7510.9	0.22	2.32	5220.0
			0.01, 1.09	2.28, 5.15	6914.7, 8166.3	0.01, 0.78	1.65, 3.38	4686.9, 5798.1
Some	60	50	1.26	4.76	−247.5	1.11	3.62	−114.8
			0.06, 4.02	2.23, 9.72	−369.2, −120.8	0.14, 3.68	1.55, 7.93	−222.4, −1.7
	60	100	2.24	9.05	−616.5	2.19	6.43	−577.2
			0.10, 6.82	4.38, 17.66	−761.6,−466.1	0.13, 6.32	2.89, 12.79	−725.4, −435.7
	250	50	1.29	5.06	−3408.1	1.27	3.89	−1746.6
			0.07, 4.01	2.38, 10.18	−4038.3, −2847.2	0.07, 3.76	1.57, 8.82	−2318.2, −1267.1
	250	100	2.16	9.00	−8766.5	2.06	6.67	-5253.5
			0.12, 6.43	4.32, 17.37	−9384.2,−8059.1	0.12, 6.04	2.94, 13.51	−6037.7, −4541.8
Yes	60	50	1.37	5.87	−369.8	1.14	3.58	−220.1
			0.07, 4.48	2.83, 13.65	−489.9, −232.8	0.05, 3.96	1.78, 8.14	−337.0, −94.7
	60	100	2.13	10.78	−638.1	1.87	6.34	−742.1
			0.11, 7.46	5.08, 23.50	−781.0, −489.0	0.08, 5.75	3.11, 13.87	−885.7, −595.7
	250	50	1.26	5.75	−5011.0	1.13	3.86	−2844.6
			0.06, 4.41	2.90, 13.15	−5642.5, −4315.1	0.04, 3.73	1.85, 9.38	−3578.0, −2168.4
	250	100	2.32	10.96	−10600.7	1.97	6.65	−8541.8
			0.10, 7.40	5.44, 24.87	−11276.8, −9890.2	0.23, 4.90	3.27, 14.80	−9431.2, −7680.7

Table 2. The minimum PBF and maximum EVD and their empirical probabilities (P) for twenty autosomal candidate regions suggested by Sabeti *et al*. (2007) as being under natural selection, based on windows of thirty SNPs

Chr:position (Mb, Hg17)	Selected population	Size (Mb)	Genes	max EVD (P)	min PBF (P)
chr1:166	CHB-JPT	0.4	*BLZF1, SLC19A2*	16.4 (0.048)	−240.7 (0.014)
chr2:72.6	CHB-JPT	0.8		17.4 (0.066)	−343.1 (0.001)
chr2:108.7	CHB-JPT	1.0	*EDAR*	19.1 (0.025)	−201.8 (0.098)
chr2:136.1	CEU	2.4	*RAB3GAP1, R3HDM1, LCT*	17.4 (0.047)	−234.4 (0.001)
chr2:177.9	CEU,CHB-JPT	1.2	*PDE11A*	23.3 (0.006)	−402.2 (0.002)
chr4:33.9	CEU,YRI,CHB-JPT	1.7		23.9 (0.002)	−170.7 (0.016)
chr4:42	CHB-JPT	0.3	*SLC30A9*	20.3 (0.022)	−419.8 (0.001)
chr4:159	CHB-JPT	0.3		17.9 (0.052)	−246.7 (0.039)
chr10:3	CEU	0.3		16.7 (0.076)	−194.8 (0.074)
chr10:22.7	CEU, CHB-JPT	0.3		19.6 (0.029)	−406.8 (0.001)
chr10:55.7	CHB-JPT	0.4	*PCDH15*	21.0 (0.020)	−206.9 (0.060)
chr12:78.3	YRI	0.8		21.9 (0.004)	−65.8 (0.124)
chr15:46.4	CEU	0.6	*SLC24A5*	25.4 (0.001)	−142.7 (0.036)
chr15:61.8	CHB-JPT	0.2	*HERC1*	20.1 (0.017)	−399.5 (0.003)
chr16:64.3	CHB-JPT	0.4		17.4 (0.025)	−321.4 (0.020)
chr16:74.3	CHB-JPT, YRI	0.6	*CHST5, ADAT1, KARS*	18.5 (0.014)	−23.8 (0.034)
chr17:53.3	CHB-JPT	0.2		13.9 (0.095)	−354.3 (0.004)
chr17:56.4	CEU	0.4	*BCAS3*	23.8 (0.003)	−344.0 (0.005)
chr19:43.5	YRI	0.3		10.7 (0.155)	−142.4 (0.130)
chr22:32.5	YRI	0.4	*LARGE*	19.1 (0.008)	−210.2 (0.041)

increasing sample size and numbers of SNPs genotyped in that region. The haplotype similarity metric (HAPS) is greatest where there is a difference between populations, as the number of SNPs increase, and when there is greater LD. Because the EVD metric is based on summary statistics, it should be less dependent on sample size. However, pairwise LD metrics are more robust with larger sample sizes (Teo *et al*., 2009). Interestingly, for the lesser sample size, the EVD credibility regions overlap for the no difference and completely different scenarios. In general, the EVD approach could be applied to windows of SNPs, and the data may be normalized using the mean and SD of the windows in that region (Teo *et al*., 2009). Irrespective of whether there are LD differences we observed a negative correlation between the PBF and EVD values (median: high LD −0.11, low LD −0.13), and PBF and HAPS absolute values (median: high LD −0.02, low LD −0.03). We found our results robust to minor changes in the specification of the prior distributions, particularly when altering the scale parameter d (values $p \pm 0.1p$). Overall, our results suggest that the PBF appears competitive with EVD and HAPS methods, with the advantages that being fully probabilistic it can be combined directly with other prior information and is directly comparable across genetic regions.

4 HAPMAP AND GAMBIAN SNP DATA

We now explore the use of the PBF on real genetic data. Using HapMap Phase II genotypic data (The International HapMap Consortium, 2007), we calculated PBF values between pairwise groups of individuals from: (i) Utah with European ancestry (CEU parents, $n = 60$), (ii) Han Chinese from Beijing (CHB) and Japanese from Tokyo (JPT) (ASIA, $n = 90$), and (iii) Yoruba people from the Ibadan region in Nigeria (YRI parents, $n = 60$). We also compared these populations to a set of controls from the self-reported Jola ethnic group ($n = 90$) from a Gambian case–control study of severe malaria (Jallow *et al*., 2009), genotyped on the Affymetrix 500K chip. Our approach captured some genes that have previously been shown to exhibit strong signatures of positive selection. We investigated 20 autosomal candidate regions that have previously been highlighted as being under natural selection (Sabeti *et al*., 2007). Table 2 summarizes the minimum PBF (and maximum EVD) values for these 20 autosomal candidate regions. In 15 regions, we

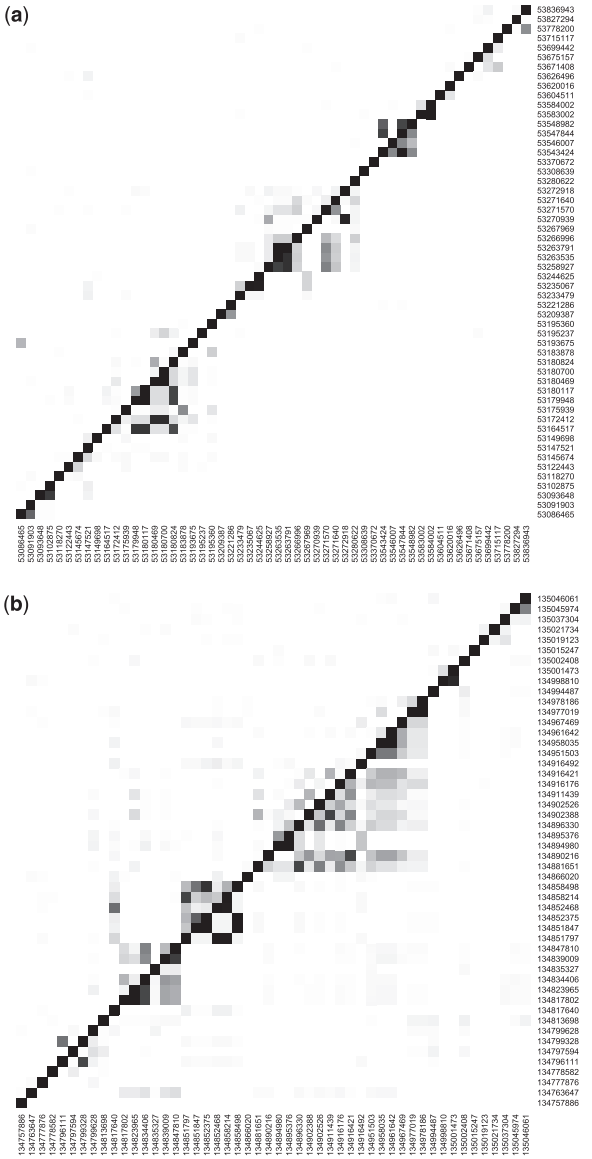

Fig. 1. **(a)** pairwise r^2 matrix for ASIA (upper) and YRI (lower) data from chromosome 153086–53837 kb; **(b)** pairwise r^2 matrix for CEU (upper) and YRI (lower) data from chromosome 2 134 758–135 046 kb; r^2 values are presented on a gray scale (0 = white, 1 = black).

observe evidence ($P < 0.05$) of an LD difference using PBF; where the P-value is calculated by permuting the data under the null and recording the distribution of resulting PBF statistics. Our results are broadly similar to the EVD approach, though we note that one of the regions of strongest selection (chr 17: 53.3 Mb) as detected by PBF was not declared as significant using EVD ($P = 0.095$). In Figure 1a, we plot the LD across ASIA and YRI populations across this candidate region on chromosome 17 (53086–53837 kb). The PBF has detected a clear extended block of LD within YRI relative to the ASIA samples. In Figure 1b, we show the LD plot for a region known to be under strong selection (chr 2: 136.1 Mb) containing the LCT gene, and was identified by the PBF as being significant ($P = 0.001$).

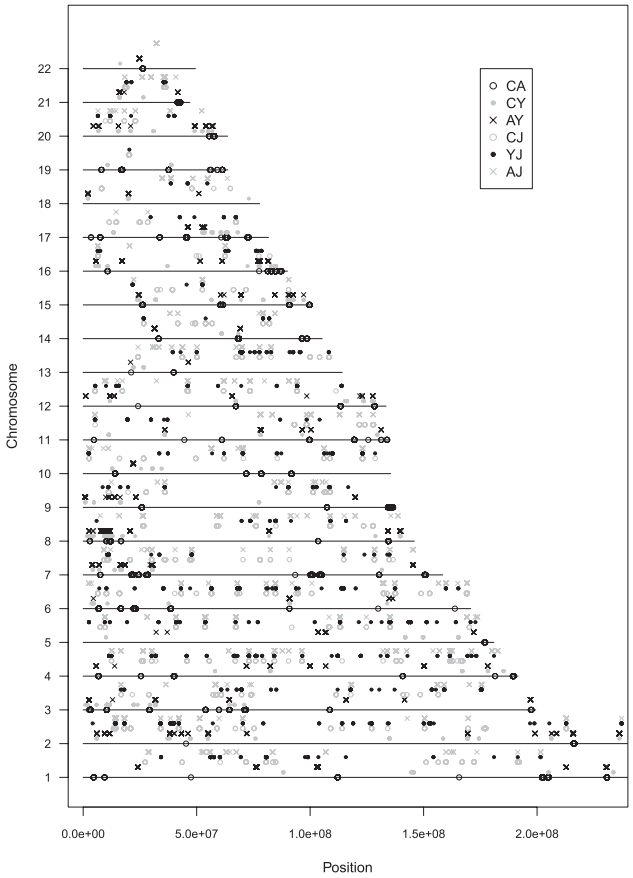

Fig. 2. The lowest 0.5% of PBF values between pairwise comparison of populations across the genome. The mean (SD) PBFs for CEU versus ASIA (CA) 566.5 (228.5), ASIA versus YRI (AY) 18.4 (222.0), CEU versus YRI (CY) 118.9 (180.0), CEU versus Jola (CJ) −933.7 (180.7), ASIA versus Jola (AJ) −1271.9 (238.9) and YRI versus Jola (YJ) −868.1 (185.4).

We then investigated the locations of the lowest 0.5% of PBF values for each pairwise population comparison, resulting in 1028 regions (192 overlapping between population comparisons, median size 483 kb, 3812 genes) (Figure 2). The majority of hits are enriched for genes related with immune response, reproduction, olfaction, morphology and metabolism (using gene ontological categories), consistent with the results from others studies (Sabeti *et al*., 2007). These include the Duffy antigen/chemokine receptor (chr 1, CEU versus YRI / Jola) (Figure 3a) and the β-globin (*HBB*) locus (chr 11, Jola versus CEU / YRI) (Figure 3b showing the large deviation in PBF across a wider region), where both regions are known to have been selected due to the advantage they provide against malaria. Similarly, we detected the *FCGR2*B and *FCGR3*B receptors (chr 1, Jola versus CEU / ASIA) and *GYPC* gene (chr 2, Jola versus CEU / YRI), which provide some protection against *Plasmodium falciparum* infection. We detected differences in the genes involved in skin pigmentation, such as the *OCA2* (chr 15, CEU versus ASIA) and *SLC24A5* (chr 15, CEU versus Jola). Other regions to be identified included the *LARGE* gene, various immune related genes (e.g. *IRF5, IL10, TGFB1, NOD1, Nostrin, CD74*), and several apolipoproteins and members of the *CYP450* family. These positive control findings support the claim the PBF is picking up real differences in LD patterns.

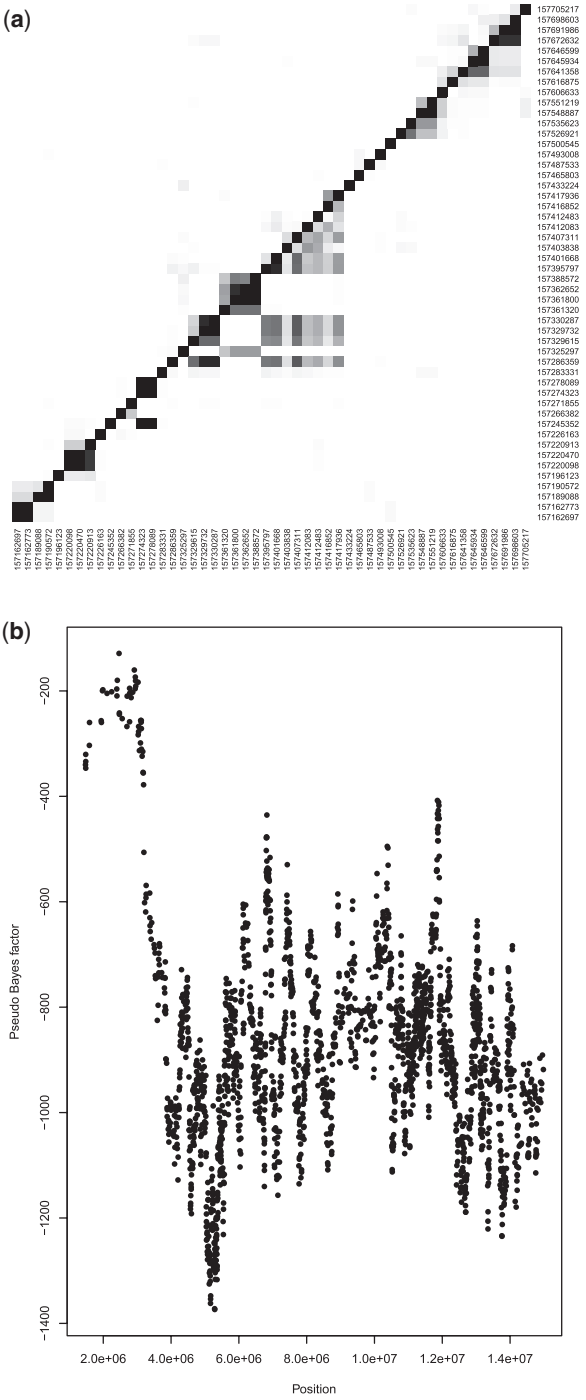

Fig. 3. (a) r^2 matrix of CEU (upper) and YRI (lower) for chromosome 1157162–157710 kb including the Duffy antigen and **(b)** plot of the PBF across a 15 Mb region of chromosome 11, where the deepest trough contains the *HBB* gene and olfactory receptors.

5 DISCUSSION

The PBF metric is a potentially robust and flexible measure to quantify differences in LD between sub-groups, which does not require time-consuming permutation approaches to infer statistical significance. Moreover, being Bayesian it readily allows for the incorporation of prior information such as knowledge of candidate regions of selection. Depending on the sub-groups of interest, regional differences in LD may be indicative of selection, recombination and/or phenotypic variation. The magnitude of the PBF reflects the difference in the covariances of genotype frequencies, accounting explicitly for the number of samples and SNPs contributing. The assumption of Gaussian allele counts is more likely to hold in non-ascertained polymorphisms and large numbers of SNPs, but our simulation and real data results suggest that method may be robust to deviations. If another distribution is considered more suitable for allele counts or non-conjugate priors adopted, then it may be possible to obtain the posterior distribution and perform inference using Markov chain Monte Carlo methods (Robert and Casella, 2004). Bayes factors can be estimated in such settings (see Clark *et al*., 2007, for an example), usually at greater computational expense. A number of other non-Bayesian approaches to compare covariance matrices could be considered as alternatives (Morrison, 1990), some robust to missing data (Jamshidian and Schott, 2007). EVD-based methods are easy to calculate, impose few assumptions, and it is possible to take advantage of known distributional properties associated with eigenvalues (Jiang, 2004; Johnstone, 2001). In addition, it is possible to extend EVD approaches to a Bayesian setting, facilitating full probabilistic inference. In conclusion, we present a new approach for quantifying LD differences in genome-wide studies, which may have potential utility in a trans-ethnic fine-mapping setting.

ACKNOWLEDGEMENTS

We thank Bronwyn MacInnis and Gareth Maslen for commenting on an earlier version of this manuscript. The study used data generated by MalariaGEN. Full lists of the investigators who contributed to the generation of these data are available at http://MalariaGEN.net. We acknowledge support from the Wellcome Trust and from the Bill and Melinda Gates Foundation, through the Foundation for the National Institutes of Health as part of the Grand Challenges in Global Health initiative, as well as from the Medical Research Council, UK.

Conflict of Interest: none declared.

REFERENCES

Bourgain,C. *et al*. (2001) Maximum identity length contrast: a powerful method for susceptibility gene detection in isolated populations. *Eur. J. Hum. Genet.*, **21** (Suppl. 1), 560–564.

Clark,T.G. *et al.* (2007) Bayesian logistic regression using a perfect phylogeny. *Biostatistics*, **8**, 32–52.

Gelman,A. *et al*. (2003) *Bayesian data analysis*. 2nd edn. Chapman and Hall, London.

Hudson,R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Jallow,M. *et al*. (2009) Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat. Genet.*, **41**, 657–665.

Jamshidian,M. and Schott,J.R. (2007) Testing equality of covariance matrices when data are incomplete. *Comput. Stat. Data Anal.*, **51**, 4227–4239.

Jiang,T. (2004) The limiting distributions of eigenvalues of sample correlation matrices. *Sankhya*, **66**, 35–48.

Johnstone,I. (2001) On the distribution of the largest eigenvalue in principal component analysis. *Ann. Stat.*, **29**, 295–327.

Kass,R.E. and Raftery,A. (1995) Bayes factors. *J. Am. Stat. Assoc.*, **90**, 773–795.

Krzanowski,W. (1993) Permutational tests for correlation matrices. *Stat. Comput.*, **3**, 37–44.

McKenzie,C.K. *et al*. (2001) Trans-ethnic fine mapping of a quantitative trait locus for circulating angiotensin i-converting enzyme (ACE). *Hum. Mol. Genet.*, **10**, 1077–1084.

Morrison,D. (1990) *Multivariate statistical methods*. McGraw–Hill,New York, pp. 292.

Robert,C.P. and Casella,G. (2004) *Monte Carlo Statistical Methods*. Springer, New York.

Sabeti,P.C. *et al*. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.

Te Meerman,G.J. and Van der Meulen,M.A. (1997) Genomic sharing surrounding alleles identical by descent: effects of genetic drift and population growth. *Genet. Epidemiol.*, **14**, 1125–1130.

Teo,Y.Y. *et al*. (2009) Genome-wide comparisons of variation in linkage disequilibrium. *Genome Res.*, **19**, 1849–1860.

Teo,Y.Y. *et al.* (2008) Power consequences of linkage disequilibrium variation between populations. *Genet. Epidemiol.*, **33**, 128–135.

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Thomas,D.C. *et al*. (2003) Bayesian spatial modeling of haplotype associations. *Hum. Hered.*, **56**, 32–40.

Tzeng,J.Y. *et al*. (2003) On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am. J. Hum. Genet.*, **72**, 891–902.

Wang,Y. *et al*. (2006) A fine-scale linkage-disequilibrium measure based on length of haplotype sharing. *Am. J. Hum. Genet.*, **78**, 615–628.

Zaykin,D.V. *et al*. (2006) Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. *Am. J. Hum. Genet.*, 78, 737–746.