In silico analysis of Single Nucleotide Polymorphisms (SNPs) in the Heparin-Binding EGF-like Growth Factor (*HBEGF*) gene and their allelic profiles in the Sri Lankan population: a comprehensive approach to prioritise SNPs for candidate gene studies

G. Gayani Harendra BSc, MSc
PhD Student, Human Genetics Unit, Faculty of Medicine, University of Colombo, Sri Lanka
E-mail address: gayanihg@gmail.com

Prof. Rohan W. Jayasekara MBBS, PhD, CBiol, MSB (Lond)
Chair and Senior Professor of Anatomy; Director and Medical Geneticist, Human Genetics Unit, Faculty of Medicine, University of Colombo, Sri Lanka
E-mail address: rohanwj@hotmail.com

Prof. Vajira H.W. Dissanayake MBBS, PhD
Professor, Department of Anatomy; Medical Geneticist, Human Genetics Unit, Faculty of Medicine, University of Colombo, Sri Lanka
E-mail address: vajirahwd@hotmail.com

Sri Lanka Journal of Bio-Medical Informatics 2011;2(1):21-37
DOI: http://dx.doi.org/10.4038/sljbmi.v2i1.2926

Abstract

In this paper, using the Heparin-Binding EGF-like growth factor (*HBEGF*) gene, we illustrate a comprehensive approach to select the most appropriate SNP markers for molecular epidemiological studies. Initially an *in silico* functional analysis was undertaken to verify the SNPs that regulate *HBEGF* expression. Thereafter based on predefined criteria (the significance of the function identified, ability to represent other SNPs in the gene (being a tagSNP), presence within an evolutionary conserved region, validation status of the SNP, and the minor allele frequency of the SNP) SNPs with putative functional effects were prioritised to identify the most appropriate *HBEGF* markers for molecular epidemiological studies. Using 30 Sinhalese men and women, we further established the allele and genotype frequencies of the seven highest priority SNPs identified. These frequencies were compared with those of HapMap populations to understand the genetic identity of the Sinhalese in relation to HapMap populations.

Keywords - In silico analysis; Single Nucleotide Polymorphisms; Heparin-Binding EGF-like growth factor; HBEGF; allelic profiles; Sri Lankan population; candidate gene studies

Introduction

Heparin-binding EGF-like Growth Factor (HBEGF) is a heparin-binding member of the EGF family[1]. It is a potent mitogen and a chemotactic factor for fibroblasts and smooth muscle cells[2] and has been shown to stimulate the growth of a variety of cells in an autocrine or paracrine manner and to be involved in stromal proliferation[3]. In addition it has the unique property of acting as the diphtheria toxin receptor[4]. Its over-expression is observed in many tumors, including hepatocarcinoma, colon, melanoma, myeloma, breast, prostate, and bladder tumors[5]. The involvement of HBEGF in various other diseases including epidermal hyperplasia[6], necrotising enterocolitis[7], cardiac hypertrophy[8] bacterial cystitis[9] and preeclampsia[10] has also been reported.

The *HBEGF* gene is composed of 13,761 bases. It lies between the region 139,692,613bp and 139,706,358bp of chromosome no. 5 (RefSeqGene Build 37.1: NC_000005.9). *HBEGF* promoter activity, as measured using a reporter gene assays, is associated with a 2.0 kb fragment upstream of the *HBEGF* gene[11]. The DNA immediately flanking the transcription start site lacks the conventional TATA and CCAAT sequences. The gene contains six exons and five introns and is consistent with the overall gene structure of the other members of the EGF family[11]. It produces a 208 amino acid residue transmembrane protein (NP_001936.1) which is encoded via a 2381bp long mRNA (NM_001945.2). The sequence contributing to the amino acid open reading frame (ORF) is found on exons 1 to 5. The dbSNP database which serves as the central, public repository for genetic variations contained a total of 128 single nucleotide polymorphisms (SNPs) in the *HBEGF* gene of *Homo sapiens* as of 15 August 2010. Functional significance has not been established for the majority of them.

In the absence of any experimental information on their functional effects, the potential functional consequences of a SNP can be predicted using various bioinformatics tools - a process known as "*in silico*"[12]. These tools predict the potential functional effects of SNPs at four main levels: splicing, transcriptional, translational and post-translational. The majority of current bioinformatics tools examine the functional effects of SNPs only with respect to a single biological function. When using such individual tools much time and effort is required to analyse SNPs and interpret the (often conflicting) predictions. There are however, few tools which provide a comprehensive assessment of SNP function. Such tools assess the deleterious effects of SNPs along functional categories, by integrating multiple tools that are based on different algorithms, data and resources[12].

We analysed all the SNPs present in the *HBEGF* gene using various composite and singleton tools to verify their putative functional effects. Those SNPs that we identified as having functional effects were then prioritised based on a number of criteria; ie. the significance of the function identified, ability to represent other SNPs in the gene (being a tagSNP), presence within an evolutionary conserved region, validation status of the SNP, and the minor allele frequency of the SNP. Accordingly, seven SNPs were selected and genotyping was carried out in 30 Sinhalese men and women to determine whether they were polymorphic. The allele and genotype frequencies obtained were then compared with those of HapMap populations to understand the genetic identity of Sinhalese.

Methodology

In silico analysis of functional effects

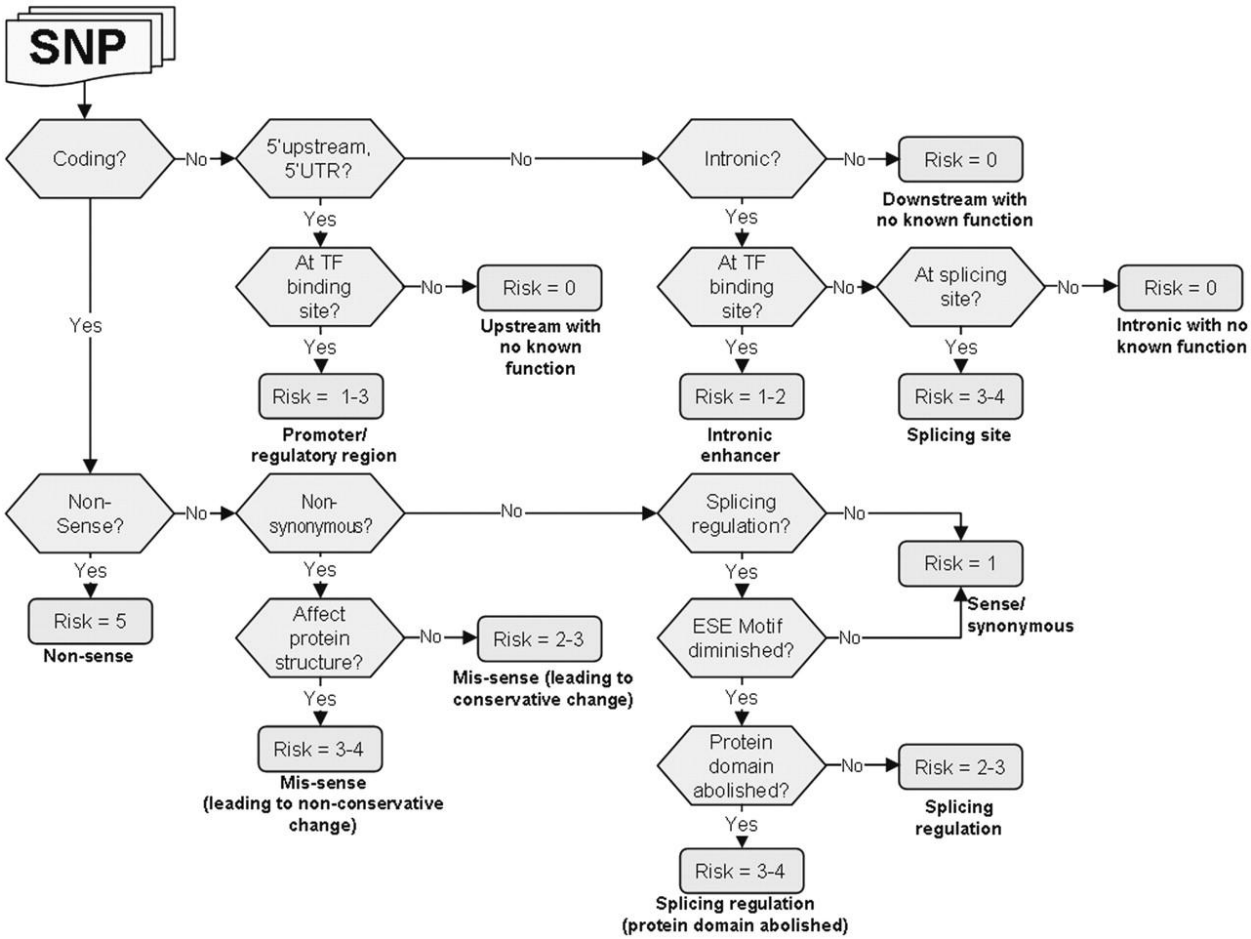
The putative functional effects were determined in both coding and non-coding regions of *HBEGF* gene as described below. In case of coding regions the effect on the protein structure was considered. To identify the effect of SNPs in non-coding regions the following features were used: Transcription factor binding sites (TFBS), Intron/exon border consensus sequences (splice sites), Exonic splicing enhancers (ESEs), and Triplex-forming oligonucleotide (TFO) target sequences.

Functional analysis was done using MATCH™, Human Splice Finder (HSF), GeneCards® database, PupaSuite2 and FASTSNP (function analysis and selection tool for single nucleotide polymorphisms). MATCH™ analyses only TFBSs. HSF analyses splice sites branch sites and ESEs. PupaSuite2 and FASTSNPs provide a comprehensive collection of functional information about SNPs using a large variety of publicly available tools and

resources. GeneCards on the other hand is an integrated database of human genes that includes automatically-mined genomic, proteomic and transcriptomic information as well as disease relationships, SNPs, gene expression and gene function.

FASTSNP (freely available at http://fastsnp.ibms. sinica.edu.tw) was used as the primary tool for SNP prioritisation and the other software were used as additional complementary tools to gain a comprehensive understanding of function. A unique feature of FASTSNP is that the prediction of functional effects is always based on the most up-to-date information which FASTSNP extracts from 11 external web servers. It prioritises SNPs according to 13 phenotypic risks and putative functional effects based on which a decision tree is prepared (Figure 1). Each effect is assigned a risk ranking - a number between 0 and 5. A high risk rank implies a high-risk level[13].

Figure 1. FASTSNP decision tree for prioritising a SNP based on its functional effects. The diamonds represent decision points and the ovals represent terminal points with the risk and class assignments. Given an input SNP at a decision point, if the answer to the question in the diamond is 'yes,' then the vertical arrow should be followed. Otherwise, the horizontal arrow should be followed. (Adapted from FASTSNP website; http://fastsnp.ibms. sinica.edu.tw).



To complement SNP prioritisation process the gene was further analysed using PupaSuite2, which is freely available at http://pupasuite.bioinfo.cipf.es. It is an interactive web-based SNP analysis tool that enables selection of relevant SNPs within a gene, based on different characteristics of the SNP itself, such as validation status, type, frequency/population data and putative functional properties (pathological SNPs, SNPs disrupting potential transcription

factor binding sites, intron/exon boundaries, exon splicing enhancers, TFO target sequences)[14]. PupaSuite2 however, does not prioritise the functional SNPs as does FASTSNP, but only searches the database to identify the functionally significant ones.

In addition to the use of these composite bioinformatic tools, all the sequence motifs containing SNPs in *HBEGF* gene were analysed for the presence of TFBS using MATCH™ software (http://www.gene-regulation.com/pub/programs.html#match). MATCH™ utilises a matrix library derived from matrices collected in TRANSFAC® and therefore provides the possibility to search for a great variety of different TFBSs[15]. Among the possible options in the tool, current analysis was conducted using the minFP option which minimises the sum of both false negative and false positive rates. In addition, using the matrix selection option, matrices were narrowed down to search only among the likely motifs found in vertebrates.

HSF (http://www.umd.be/HSF/) was used to identify possible splice sites, branch sites and other *cis* acting elements in the *HBEGF* gene. The programme utilises a new algorithm derived from the Universal Mutation Database (UMD) to evaluate the splice sites and branch points. In addition to new algorithms, it includes already published algorithms as well, such as RESCUE-ESE and ESE-Finder which are made use of in identifying *cis*-acting elements[16]. Since our major interest was on the effect of SNPs in creating or abolishing regulatory sites we used the 'analyse mutations' option. The threshold value was set at 80 for creation or abolition of splice sites. The threshold value was set at 75 for predicting ESE.

The results obtained with the above tools were combined with the information available at GeneCards (http://www.genecards.org/) to get a more complete view of gene regulation. In contrast to other bioinformatics tools which analyse the gene sequences based on given algorithm(s) to locate specific set of sequence motifs, GeneCards act as an all-inclusive database which integrate the fragments of information scattered over a variety of specialised databases into a coherent picture. This is especially important to locate ESEs sitting on SNPs as HSF does not have the facility to analyse most untranslated gene regions.

In addition to screening for the putative functional effects, linkage disequilibrium (LD) among the SNPs and the SNP distribution in relation to evolutionary conserved areas were also analysed as described below and was utilised in the SNP prioritisation process.

Selection of tag SNPs

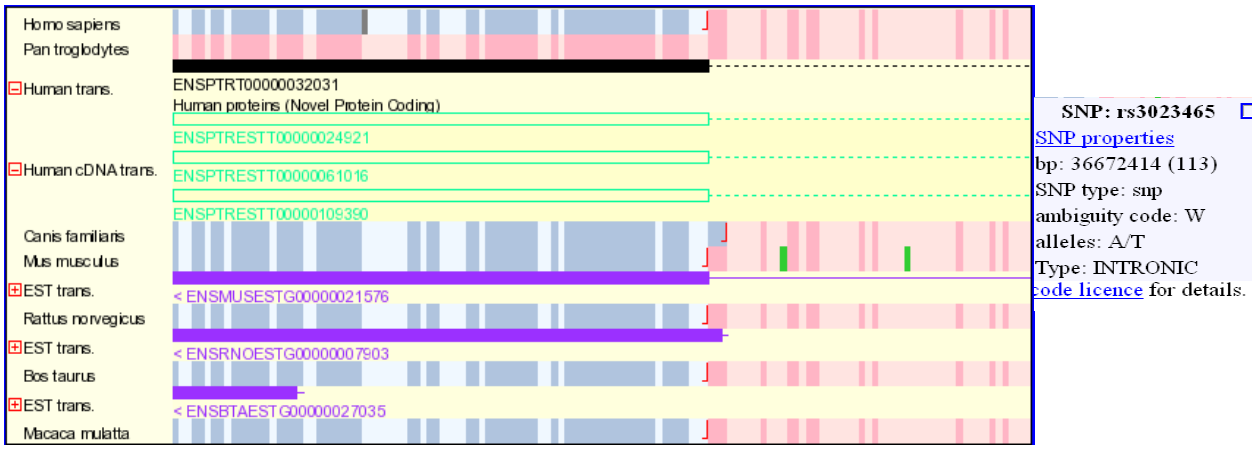
TagSNPs selection was carried out using Haploview, which allows visualising LD patterns in genetic data (http://www.broad.mit.edu/mpg/haploview/). It directly accepts genotype data downloaded from the Human HapMap website (http://www.hapmap.org/) for analysis. Therefore, HapMap data available for the *HBEGF* gene including a 2kb area flanking the gene on 5'end and 600bp area from 3' end was selected for tagSNP analysis. Phase 2 genotype data from the Han Chinese population in Beijing (CHB) was selected and LD analysis was carried out using the default algorithm. Based on the LD pattern, tagSNP selection was carried out with the pairwise tagging method. For this analysis r^2 threshold was set at 0.8 without forced inclusion or exclusion of any SNPs.

Identifying SNPs located in evolutionary conserved regions in the gene

SNPs located in evolutionary conserved regions (cSNPs) were identified using the Ensemble Genome browser - release 48 (http://www.ensembl.org/). Six eutherian mammals were

selected for the comparative analysis with the human *HBEGF* gene; namely chimpanzee (*Pan troglodytes*), dog (*Canis familiaris*), rat (*Rattus norvegicus*), macaque (*Macaca mulatta*), cow (*Bos taurus*) and mouse (*Mus_musculus*). Alignments with 7 eutherian mammals Pecan under *Ensembl Human AlignSlice View* was used to retrieve the comparative genomic information. For a detailed comparison, base pair view was selected from the database and the SNPs were manually checked to verify whether they fall within the conserved region identified by the browser. Figure 2 shows a screenshot of the base pair view which compares the genomic region 139702500bp to 139702630bp of human chromosome no. 5 with seven selected eutherian mammals.

Figure 2. A comparison of the human *HBEGF* gene with six selected eutherian mammals: the screenshot compares human chromosome no. 5:139702500 bp to 139702630 bp region with homologous regions of other mammals. Exons are represented by the blue background while all other nucleotides are represented by the pink background. Red marks demarcate the exon boundaries. Shades of blue and pink represent the similarity level of the species: darker colours represent highly conserved areas of nucleotides. Gray and green marks represent SNPs. A description of a SNP that can be obtained by selecting a gray or green line is also shown. [The screen shot was taken from Ensemble Genome browser- release 48 (http://www.ensembl.org/)].



SNP prioritsation for genotyping

The SNPs identified as having putative functional effects, cSNPs and TagSNPs were then screened for their minor allele frequency (MAF) and their validation status. SNPs that are not validated or that have a MAF ≤0.05 were not considered for genotyping except rs41445951. rs41445951 was force included in the SNP list for genotyping, despite its validation status and MAF, since it was the only nonsynonymous SNP reported in the *HBEGF* gene. Among the SNPs having a functional effect, the tag SNPs, cSNPs and those that had a MAF ≥ 0.10 were given a higher priority. Accordingly seven SNPs (rs41445951, rs2074611, rs1862176, rs4150196, rs13385, rs2237076 and rs2074613) were selected for genotyping.

SNP Genotyping

Genotyping was conducted using an already established DNA resource at the Human Genetics Unit, Faculty of Medicine, Colombo. This collection had been made for studies of this nature according to protocols approved by the Ethics Review Committee of the Faculty of Medicine, University of Colombo. All subjects were recruited for these studies in

Colombo, Sri Lanka between August 2001 and January 2003. All subjects had given written informed consent to participate. The study was carried out in accordance with the Declaration of Helsinki of the World Medical Association. 30 samples of Sinhalese men and women from this resource were randomly selected for this study. The ethnicity of these subjects had been established at the time of collection by inquiring into the grandparental ethnicity.

rs41445951 and rs2074611 were genotyped using newly designed PCR-RFLP (polymerase chain reaction restriction fragment length polymorphism) assays while rs1862176, rs4150196, rs13385, rs2237076 and rs2074613 were genotyped using MassArray Sequenom iplex methodology. The details of the genotyping methods are available on request.

Statistical analysis

The chi-squared test was used to test the genotypes at each polymorphic locus for Hardy–Weinberg Equilibrium (HWE). Estimation of haplotype frequencies and measurement of pairwise linkage disequilibrium (D' and r2) were carried out using Haploview[17].

Results

Functional effects of SNPs

Putative functional effects of the SNPs that lie within *HBEGF* gene were analysed using numerous bioinformatic tools; viz. FASTSNP, PupaSuite2, HSF, MATCH™, Ensemble AlignSlice view and GeneCards® database. The results of this analysis are in Table 1. Risk factors were assigned according to the criteria used in FASTSNP (see Figure 1). SNPs that affect TFO target sequences were not assigned a risk factor as this is not included in the FASTSNP criteria. As shown in the table, out of a total of 128 SNPs reported in the *HBEGF* gene, 63 SNPs had a putative functional effect. The highest risk was observed for those that reside within the exonic (11 SNPs) region (coding and UTR; risk factor 2-3). Eleven more SNPs residing within the 5'upstream region were identified as having relatively low risk of 1-3 by interfering promoter function. The other 38 SNPs that were identified to have putative functional effect disrupt intronic enhancers and were categorised as having low risk of 1-2.

Among the identified functional SNPs, 14 were found to be cSNPs. Another SNP, which did not show any putative functional effect, was also found to be in a conserved region. This observation, i.e. that a majority of cSNPs have some functional effect, supports the general view which stresses the importance of focusing on evolutionary conserved areas of the genome in predicting functionally significant markers.

TagSNP selection

TagSNP analysis was carried out based on the genotype data available from the HapMap Project for 45 individuals from a Han Chinese population in Beijing, China (HCB). For them, genotype data for twenty seven SNPs were available for the *HBEGF* gene. Out of these, only 11 SNPs had a MAF $\geq$0.05 and were considered for LD analysis. The rest were rejected by the default algorithm of Haploview due to their very low minor allele frequency.

Table 1. A summary of SNP function prediction obtained with various bioinformatic tools and their validation status.

SNP ID	Possible Functional Effects	Risk	Gene Region	Found in conserved region	MAF in Asians
rs2074611§	Sense/synonymous; SR	2-3	coding		0.07
rs41445951	Missense (conservative) /SR	2-3	coding		0.02
rs4150241§	SR	2-3	3'UTR		NA
rs4150240§	SR	2-3	3'UTR		NA
rs13385§	SR*	2-3	3'UTR		0.2
rs4150239§	SR	2-3	3'UTR		0
rs11465459§	SR	2-3	3'UTR	cSNP	0
rs12656477§	SR	2-3	3'UTR	cSNP	0
rs4150238§	SR	2-3	3'UTR	cSNP	0.01
rs4150237§	SR	2-3	3'UTR		0.1250
rs1042184§	SR	2-3	5'UTR	cSNP	0
rs11465434§	P/R	1-3	5'upstream	cSNP	NA
rs4150196§	P/R	1-3	5'upstream		0.841
rs3776089§	P/R	1-3	5'upstream		0.133
rs4150195§	P/R	1-3	5'upstream		0
rs4150193§	P/R	1-3	5'upstream		0
rs11465432§	P/R	1-3	5'upstream		0
rs11465431§	P/R	1-3	5'upstream		0
rs6890393§:	P/R	1-3	5'upstream		0.0
rs6879095§	P/R	1-3	5'upstream		NA
rs1862176§:	P/R	1-3	5'upstream		0.133
rs57817790	P/R	1-3	5'upstream		N/A
rs2074613§	IE	1-2	intronic		0.42
rs4150235§	IE	1-2	intronic		NA
rs41291441	IE	1-2	intronic		NA
rs6882074§	IE	1-2	intronic		NA
rs4150228§	IE	1-2	intronic		NA
rs4150227§	IE	1-2	intronic		0
rs4150226§	IE	1-2	intronic		0
rs11465448§	IE	1-2	intronic		NA
rs35630316§	IE	1-2	intronic		NA
rs11465447	IE	1-2	intronic		NA
rs73273161	IE	1-2	intronic		NA
rs11168045§	IE	1-2	intronic		NA
rs4150221§	IE	1-2	intronic		NA
rs35411477	IE	1-2	intronic		NA
rs35688900	IE	1-2	intronic		NA
rs58992612§	IE	1-2	intronic		NA
rs4150220	IE	1-2	intronic		NA
rs2237076§	IE	1-2	intronic		0.16
rs11465445§	IE	1-2	intronic		NA
rs2237075§	IE	1-2	intronic		0.156
rs4150215§	IE	1-2	intronic		NA
rs4150214§	IE	1-2	intronic		0.125
rs11465439§	IE	1-2	intronic		NA
rs4150212§	IE	1-2	intronic		0.133
rs11465438§	IE	1-2	intronic		NA
rs4150210§	IE	1-2	intronic		0.125
rs4150209§	IE	1-2	intronic		0
rs4150207§	IE	1-2	intronic		NA
rs4150206§	IE	1-2	intronic		NA
rs2282801	IE	1-2	intronic		NA

SNP ID	Possible Functional Effects	Risk	Gene Region	Found in conserved region	MAF in Asians
rs4150205[§]	IE	1-2	intronic		NA
rs4150204	IE	1-2	intronic		NA
rs11465436	IE	1-2	intronic		NA
rs11465435[§]	IE	1-2	intronic		NA
rs35507314	IE	1-2	intronic		NA
rs2237078[§]	IE	1-2	Intronic		0.140
rs35665233	IE	1-2	Intronic	cSNP	NA
rs6889944	IE	1-2	Intronic		NA
rs4150225[§]	TFO		Intronic		0
rs4150216[§]	TFO		Intronic		0.114
rs5871732	TFO		Intronic		NA
rs4150229[§]	None	0	Intronic	cSNP	NA

SR: splicing regulation; P/R: promoter/regulatory region, IE: intronic enhancers, UTR: untranslated region, MAF: minor allele frequency; cSNP: SNP in conserved region, [§] :validated SNP; * :tagSNP

The graphical representations of LD patterns and r^2 created by Haploview are shown in Figure 3. As shown two main haplotype blocks were identified for the population into which all SNPs except rs4912711 was included. Block 1 encompassed a 4.1 kb area of the gene extending from exon 6 to intron 3 (chromosome position: 139,712878 to 139,716,988) and was defined by 4 SNPs. Block 2 encompassed a 6.8 kb area of the gene extending from intron 3 to the promoter region (chromosome position: 139,720400 to 139727260) and was defined by 6 SNPs. Accordingly, the SNP density used for LD analysis was roughly around 1 SNP per 1 kb of the gene. The results obtained for TagSNP analysis based on the observed LD pattern is in Table 2.

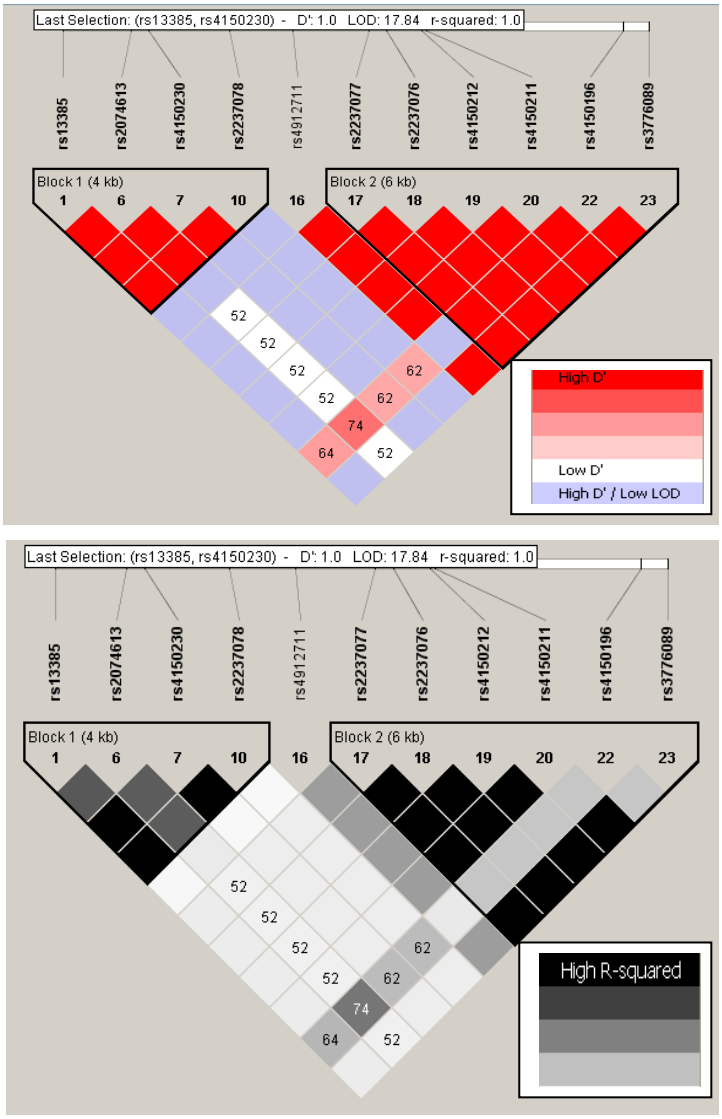
Table 2. TagSNPs identified (at r^2=0.8) for *HBEGF* gene using HapMap Project data for Han Chinese in Beijing, China (HCB)

TagSNPs	Alleles captured
rs4150212	rs4150211, rs2237076, rs2237077, rs3776089, rs4150212
rs13385	rs4150230, rs13385, rs2237078

Figure 3. Linkage disequilibrium (LD) plot (a) and R^2 plot (b) generated using the default settings in Haploview for Han Chinese in Beijing, China (HCB).

a)

b)



Note: LD between pairs of markers across HBEGF gene and near gene region was determined within Haploview using (a) D' or (b) r^2 statistics. LD values between markers are indicated at the intercept of the two markers on the matrix. Empty boxes indicate that the LD value is one. Intensity of color on the red/pink or black/gray scale indicates the degree of confidence in the LD value. The two main haplotype blocks outlined within the black triangles encompass; Block one: exon 6 to intron 3 of the *HBEGF* gene; Block two: introns 3 to promoter region of the *HBEGF* gene. All single-nucleotide polymorphisms (SNPs) used had minor allele frequencies ≥0.05.

Allele and genotype frequencies of selected SNPs

Genotype and minor allele frequencies of the seven selected SNPs are summarized in Table 3. A comparison of heterozygocity and MAF data of Sinhalese with several other populations reported in scientific literature is given in Table 4. As shown, rs41445951, the only missense SNP reported in the *HBEGF* gene, was not polymorphic among Sinhalese. In fact it was reported to be polymorphic only in the R24 population. The other six SNPs were found to be polymorphic among Sinhalese with genotype frequencies conferring to Hardy Weinberg Equilibrium (HWE). Among them, heterozygosity (H) of rs2074613 (0.47) was close to the

maximum possible heterozygosity (0.5) within the HWE conditions. This value was comparable with the average heterozygosity (AvgH) reported for the other populations (0.48). Even though rs4150196 also reached a relatively higher level of heterozygosity among Sinhalese (H = 0.40), the reported AvgH for other populations were much higher and close to the maximum H within HWE (0.498). Among the six SNPs found to be polymorphic, rs13385 had the lowest H (0.22) in both Sinhalese (0.22) and other populations (0.33) with H observed among Sinhalese being much lower than that of others. For the other SNPs (viz. rs2074611, rs1862176 and rs2237076), heterozygosities observed for Sinhalese (0.35, 0.33 and 0.37) was comparable to the AvgH reported in other populations (0.37, 0.39 and 0.39).

When the minor alleles (MA) observed for Sinhalese were compared with those of other HapMap populations, rs4150196 and rs2074613 showed significant deviations from that of HCB and CEU. In the case of rs4150196, the minor allele was A in these populations compared to G in the Sinhalese. In the case of rs2074613 it was G as opposed to A in the Sinhalese. In addition to these two populations, the YRI population also had G as the minor allele for rs2074613. Likewise for rs4150196, the E0 population, in addition to HCB and CEU, had A as the minor allele. Obviously this condition of not having an established minor allele in all populations must have been brought about by the high AvgH observed for these two SNPs. However, rs2237076, which had a relatively low AvgH (0.39) when compared to the two former SNPs, also showed a change in the minor allele in the E1 population (A) when compared with the minor alleles in other populations (G). However, JPT and D0 populations did not show any minor allele deviations compared to Sinhalese with respect to the six polymorphic SNPs genotyped.

Table 3. Summary of genotype and minor allele frequencies (MAF) observed in Sinhalese

SNP	Location	Putative functional effect	Genotype frequencies			N	MAF
rs41445951	Ex 2	Missense/splicing regulation	CC: 30 (1.0)	CT: 0	TT: 0	30	0
rs2074611	Ex 3	Sense/splicing regulation	CC: 16 (0.53)	CT: 14 (0.47)	TT: 00 (0.00)	30	0.23
rs1862176	P	Promoter/regulatory	AA: 17 (0.60)	GA: 10 (0.36)	GG: 01 (0.04)	28	0.21
rs4150196	P	Promoter/regulatory	AA: 09 (0.33)	AG: 15 (0.56)	GG: 03 (0.11)	27	0.39
rs13385	Ex 6; 3'UTR	Exonic splicing enhancer (tag SNP)	CC: 14 (0.50)	CT: 09 (0.32)	TT: 05 (0.18)	28	0.34
rs2237076	Int 3	Intronic enhancer	GG: 16 (0.57)	GA: 10 (0.36)	AA: 02 (0.07)	28	0.25
rs2074613	Int 4	Intronic enhancer	GG: 9 (0.32)	GA: 16 (0.57)	AA: 03 (0.11)	28	0.39

Genotype frequencies are given as n(f) where n=number of samples, f=frequency;
N = total number of samples genotyped successfully. UTR: untranslated region, Ex: Exon, Int: Intron, P: promotor

Table 4. A comparison of average heterozygosity (AvgH) and minor allele frequency (MAF) of Sinhalese with other populations for the seven selected SNPs.

SNP	**AvgH**		**Average MAF**							
	Sinhalese	**Others**	**Sinhalese**	**Asians**		**Europeans**			**African Americans**	**Africans**	**Global**
				HCB[†]	**JPT**[†]	**CEU**[†]	**E1**[§]	**E0**[§]	**D0**[§]	**YRI**[†]	**PRD90**[§]
rs41445951	**0**	0.04	**T: 0**	ND	ND	ND	ND	ND	ND	ND	T: 0.02[*]
rs2074611	**0.35**	0.37	**T: 0.23**	T: 0.07	T: 0.11	T: 0.05	T: 0.5	T: 0.05	T: 0.29	T: 0.23	T: 0.12
rs1862176	**0.33**	0.39	**G: 0.21**	G: 0.13	G: 0.14	G: 0.25	ND	ND	ND	G: 0.48	ND
rs4150196	**0.40**	0.49	**G: 0.39**	A: 0.41	G: 0.48	A: 0.39	G: 0.33	A: 0.45	G: 0.37	G: 0.27	G: 0.45
rs13385	**0.22**	0.35	**T: 0.34**	T: 0.32	T: 0.41	T.0.26	T: 0.50	T: 0.15	T: 0.17	T: 0.03	T: 0.18
rs2237076	**0.37**	0.39	**A: 0.25**	A: 0.13	A: 0.13	A: 0.21	G: 0.33	A: 0.20	A: 0.41	A: 0.44	A: 0.33
rs2074613	**0.47**	0.48	**A: 0.39**	G: 0.42	A: 0.39	G: 0.42	A: 0.33	ND	ND	G: 0.22	G: 0.43

[†]HapMap populations; [*]R24 population frequency, [§]non-HapMap populations for whom genotype data are available for the selected SNPs as found in NCBI (http://www.ncbi.nlm.nih.gov/SNP/) ND: data not available.

The haplotype frequencies inferred by the distribution of observed genotypes in Sinhalese is given in Table 5. As shown, two haplotypes (CAGCGA and TGGCAA) were found to be the predominant ones accounting for 45.5% of chromosomes. 40.5% of chromosomes were represented by twelve different haplotypes occurring at a frequency less than 0.05. The results of the LD analysis (Table 6) indicates a relatively high LD between rs1862176 and rs2237076 (D'=0.888, R^2=0.637).

Table 5. Frequencies of haplotypes defined by the six polymorphic loci

Haplotype						Frequency
rs13385	**rs2074613**	**rs2237076**	**rs2074611**	**rs4150196**	**rs1862176**	
C	A	G	C	G	A	0.290
T	G	G	C	A	A	0.165
C	G	A	C	A	G	0.085
C	G	G	C	A	A	0.055
†Others						0.405

† includes 12 more haplotypes which are showing frequencies less than 0.05

Table 6. Linkage disequilibrium (LD) analysis among polymorphic loci

SNP	**rs1862176**		**rs4150196**		**rs2074611**		**rs2237076**		**Rs2074613**	
	D'	**R^2**	**D'**	**R^2**	**D^2**	**R^2**	**D'**	**R^2**	**D'**	**R^2**
rs4150196	1.0	0.173								
rs2074611	1.0	0.1	0.057	0.002						
rs2237076	**0.883**	**0.637**	1.0	0.215	0.25	0.057				
rs2074613	0.475	0.04	0.704	0.455	0.048	0.001	0.595	0.076		
rs13385	0.034	0.0	0.625	0.123	0.149	0.016	0.078	0.001	1.0	0.332

D':Lewontin disequilibrium coefficient; R^2: correlation coefficient

Discussion

Due to the large number of SNPs present in the human genome it is difficult to prioritise and select candidate markers for disease association studies. While much attention has focused on variants in protein coding DNA, variants in non-coding regions may also play an important role in the aetiology of complex diseases by altering gene regulation. Since the vast majority of non-coding genomic sequence are of unknown function this presents a challenge to identify functional variants that cause diseases. However *in silico* functional analysis coupled with other appropriate information could be used as a guideline to determine the most likely causative variants. We bioinformatically analysed the *HBEGF* gene, which is implicated in various complex diseases, to identify and prioritise the functionally important SNPs present in the gene. A selected set of SNPs were then genotyped in a Sinhalese population.

It is well known that no single bioinformatic tool can be used to obtain a complete picture of the functional significance of allelic variants. Hence the current analysis was conducted using a number of complementary bioinformatic tools. As expected the results obtained from

different tools did not directly overlap. Of the composite tools used, the highest number of functional SNPs was identified using FASTSNP. FASTSNP provides a risk ranking for all the variants present in dbSNP. Some other tools like Pupasuit2 and HSF cannot be used for un-translated regions and/or non validated SNPs. On the other hand FASTSNP did not identify some of the functions identified by either MATCH^{TM} or HSF indicating the importance of incorporating a broader range of tools in the analysis.

Even when using multiple tools, it is essential to take precautions to avoid the common pitfalls that one might come across when conducting an *in silico* analysis. For example it is understood that not all SNPs in dbSNP are real. Some polymorphisms might have arisen solely due to sequencing errors and others may be unique to the individuals they were found in and may not be found in others[18]. Hence it is often possible to follow a nonexistent SNP which could be avoided by looking at the validation status, which reflects the reliability of a given SNP. The non polymorphic nature of rs41445951 in the Sinhalese population, found with the current study, is an example. Thus, it is always recommended to treat non-validated SNPs with caution.

Another common problem that one comes across with *in silico* prediction tools is the high rate of false positive findings produced by them due to the short length of sequences (typically 6–8-mer) used in computer simulations[19]. The additional information about conserved regions across multiple species could be used as a way to filter out such false-positive predictions[19-21]. Thus, SNPs that lie within an evolutionary conserved region could be considered as representing a truly functional variant in comparison to SNPs that lie outside such regions. This strategy was used in the current study to validate the *in silico* predictions. However, due to the rather low MAFs showed by most such cSNPs they could not be prioritised.

The use of MAF to prioritise functional SNPs identified with *in silico* tools is related to statistical power of the study. For instance, the power to detect variant allele of a SNP in a given sample population depends on the MAF of the polymorphism; i.e. The lower the MAF the number of samples need to detect the variant allele increases[22]. Therefore, SNPs with a MAF of 5% or more are generally targeted in the majority of large scale genome studies such as the international HapMap project. The same criteria were used for the present analysis as well and a higher priority was given to those SNPs with a MAF of 10% or more.

LD is another important consideration when selecting SNPs for genotyping. Empirical studies suggest that much of the human genome can be characterised as blocks of strong linkage disequilibrium (the non-random association of alleles at two or more loci) or haplotypes[22]. With strong correlation between markers, much of the common haplotype diversity can be represented by a small number of representative SNPs which are known as tagSNPs. For example, when two SNPs are in perfect LD (R^2=1) they are regarded redundant for genotyping. On the other hand zero LD between two SNPs indicates loss of representative power of the two SNPs in relation to each other. An R^2 of 0.8 is often used for tagSNP selection, as this value is felt to represent a compromise between completeness and efficiency of coverage of the common variation in the genome[22]. Since tagSNPs contain most of the information that could be gained by genotyping the other surrounding SNPs, using them in association studies result in a substantial reduction in genotyping costs with only a minimal loss of power. Hence, in the current study tagSNPs carrying a potential functional effect were given a higher priority compared to those functional SNPs which were not in strong LD with others.

When one consider the SNPs selected for genotyping, except rs41445951 all had a relatively high MAF for Asian populations. rs41445951 was selected disregarding the fact that it was not validated and the fact that it had a very low MAF (0.02), because it was the only reported missense SNP in the *HBEGF* gene. Thus, if it was present in our population, it would have got the highest functional significance of all the SNPs present within the gene. Among the other selected SNPs rs13385 was a tagSNP in addition to being an ESE. However, the other tagSNP (rs4150212) identified with Haploview was not selected for genotyping despite it being a putative intronic enhancer. This exclusion was done due to the fact that there was no Asian genotyping data available for this SNP to indicate the presence of this polymorphism among Sinhalese.

Associations between multilocus heterozygosity and fitness traits, also termed heterozygosity and fitness correlations (HFCs), have been reported in numerous organisms[23]. These studies, in general, indicate a positive relationship between heterozygosity and fitness traits. For example, a heterozygous advantage is observed in the expression of various disease phenotypes and loss of SNP heterozygosity is shown to be associated with cancer risk[24]. In addition heterozygosity is important in revealing population history[25]. In this respect, it is interesting to note the high heterozygosity rates observed for rs4150196, rs13385, and rs2074613, which need further investigation to reveal any possible associations with disease traits and population demography.

According to the genotyping results, it is evident that the Sinhalese population has a higher similarity with JPT than with HCB with respect to the seven selected SNPs. This was apparent with rs4150196 and rs2074613 where Sinhalese and JPT shared the same minor alleles while CHB had a different minor allele along with CEU. This knowledge of genetic similarity is important in tagSNP selection which is performed prior to selecting markers for genotyping in association studies. In the current study, due to lack of available Sinhalese genotype data to compare with other populations, HCB was chosen arbitrarily among the two available Asian HapMap populations to use as the base population for tagSNP analysis. It is highly likely that different sets of tagSNPs would be found in different populations (for the *HBEGF* gene, rs3776089 and rs4150230 becomes tagSNPs with JPT data) which would affect the SNP prioritisation process. Therefore, in addition to clarifying the Sinhalese genetic composition with respect to the seven selected SNPs this study provides a baseline that can be used to compare the Sinhalese with other populations in future studies.

Conclusion

This report exemplifies the use of bioinformatic tools in marker selection for candidate gene studies and provides a comprehensive analysis of functional effect of SNPs present in the *HBEGF* gene. In addition, we report the genetic composition of Sinhalese with respect to seven SNPs and provide a database that can be used as a guide for future genetic studies in the Sinhalese population.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Acknowledgements

This study was funded by research grants from the National Research Council of Sri Lanka (Grant No: 05-15) and the Human Genetics Unit Development Fund, University of Colombo, Sri Lanka. The PhD studentship of GGH was funded by grants from the University of Colombo and the University Grants Commission, Sri Lanka (Grant No: UGC/ICD/RPC-2008/26).

References

1. Raab G, Klagsbrun M. Heparin-binding EGF-like growth factor. *Biochim Biophys Acta* 1997; **1333**:F179-99.

2. Blotnick S, Peoples GE, Freeman MR, Eberlein TJ, Klagsbrun MT. Lymphocytes synthesize and export heparin-binding epidermal growth factor-like growth factor and basic fibroblast growth factor, mitogens for vascular cells and fibroblasts: differential production and release by CD4 and CD8 T cells. *Proc Natl Acad Sci USA* 1994; **91**:2890-4.

3. Hisaka T, Yano H, Haramaki M, Utsunomiya I, Kojiro M. Expressions of epidermal growth factor family and its receptor in hepatocellular carcinoma cell lines: relationship to cell proliferation. *Int J Oncol* 1999; **14**:453-60. PMID:10024677 [PubMed - indexed for MEDLINE].

4. Raab G, Higashiyama S, Hetelekids S, Abraham JA, D. Damm, *et al.* Biosynthesis and processing by phorbol ester of the cell surface-associated precursor form of heparin-binding EGF-like growth factor. *Biochem Biophys Res Commun* 1994; **204**:592 7. http://dx.doi.org/10.1006/bbrc.1994.2500

5. Ongusaha PP, Kwak J, Zwible AJ, Macip S, Higashiyama S, *et al.* HB-EGF is a potent inducer of tumor growth and angiogenesis. *Cancer Res* 2004; **64**:5283-90. http://dx.doi.org/10.1158/0008-5472.CAN-04-0925

6. Rittié L, Varani J, Kang S, Voorhees JJ, Fisher GJ. Retinoid-induced epidermal hyperplasia is mediated by Epidermal Growth Factor receptor activation via specific induction of its ligands Heparin-Binding EGF and Amphiregulin in human skin in vivo. *J Invest Dermatol* 2006; **126**:732-9. http://dx.doi.org/10.1038/sj.jid.5700202

7. Jiexiong F, Osama NE-A, Gail EB. Heparin-binding EGF-like growth factor (HB-EGF) and necrotizing enterocolitis. *Semin Pediatr Surg* 2005; **14**:167-74. http://dx.doi.org/10.1053/j.sempedsurg.2005.05.005

8. Seiji T, Masafumi K. HB-EGF, Transactivation, and Cardiac Hypertrophy. *Int J Gerontol* 2007; **1**:2-9.

9. Zhang C-O, Li Z-L, Kong C-Z. APF, HB-EGF, and EGF biomarkers in patients with ulcerative vs. non-ulcerative interstitial cystitis. *BMC Urol* 2005; **5**:7. http://dx.doi.org/10.1186/1471-2490-5-7

10. Leach RE, Romero R, Kim YM, Chaiworapongsa T, Kilburn B, *et al.* Pre-eclampsia and expression of heparin- binding EGF-like growth factor. *The Lancet* 2002; **360**:1215-9. http://dx.doi.org/10.1016/S0140-6736(02)11283-9

11. Fen Z, Dhadly MS, Yoshizumi M, Hilkert RJ, Quertermous T, *et al.* Structural organization and chromosomal assignment of the gene encoding the human heparin-binding epidermal growth factor-like growth factor/diphtheria toxin receptor. *Biochemistry (Mosc)* 1993; **32**:7932-8. http://dx.doi.org/10.1021/bi00082a014

12. Bhatti P, Church DM, Rutter JL, Struewing JP, Sigurdson AJ. Candidate Single Nucleotide Polymorphism Selection using Publicly Available Tools: A Guide for Epidemiologists. *Am J Epidemiol* 2006; **164**:794-804. http://dx.doi.org/10.1093/aje/kwj269

13. Yuan H, Chiou J, Tseng W, Liu C, Lin Y, Wang H, *et al.* FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Res* 2006; **34** (Web Server issue):W635-W41. http://dx.doi.org/10.1093/nar/gkl236

14. Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, *et al.* PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic Acids Res* 2006; **34** (Web Server Issue):W621-W5. http://dx.doi.org/10.1093/nar/gkl071

15. Kel AE, Gling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, *et al.* MATCH^{TM}: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* 2003 July 1, 2003; **31**:3576-9. http://dx.doi.org/10.1093/nar/gkl071

16. Desmet F-O, Hamroun D, Lalande M, Collod-Broud G, Claustres M, *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 2009; **37**:1-14. http://dx.doi.org/10.1093/nar/gkp215

17. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2004; **21**:263-5. http://dx.doi.org/10.1093/bioinformatics/bth457

18. Finn R, Forrest M, loveland J, Overduin B, Rabinowicz P, *et al.* The open door workshop-Working with the human genome sequence-Course Manual. 2009; Bangkok, Thailand. Wellcome Trust-Mahidol University-Oxford Tropical Medicine Programme; 2009. p. 201.

19. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 2004; **11**:377-94. http://dx.doi.org/10.1089/1066527041410418

20. Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. *Science* 2002; **297**:1007-13. http://dx.doi.org/10.1126/science.1073774

21. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res* 2003; **31**:3568-71.

22. Grover D, Woodfield AS, Verma R, Zandi PP, Levinson DF, *et al.* QuickSNP: an automated web server for selection of tagSNPs. *Nucleic Acids Res.* (Web Server issue) 2007; **35**:W115-W20. http://dx.doi.org/10.1093/nar/gkg616

23. David P. Heterozygosity-fitness correlations: new perspectives on old problems. *Heredity* 1998; **80**:531-7. http://dx.doi.org/10.1046/j.1365-2540.1998.00393

24. Ruivenkamp C, Hermsen M, Postma C, Klous A, Baak J, *et al.* LOH of PTPRJ occurs early in colorectal cancer and is associated with chromosomal loss of 18q12-21. *Oncogene* 2003; **22**:3472-4. http://dx.doi.org/10.1038/sj.onc.1206246

25. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution* 1984; **38**:1358-70.