Characterisation of SNP haplotype structure in chemokine and chemokine receptor genes using CEPH pedigrees and statistical estimation

Vanessa J. Clark[1,2]* *and Michael Dean*[1]

[1]Laboratory of Genomic Diversity, Human Genetics Section, National Cancer Institute, Frederick, MD 21702, USA
[2]Department of Human Genetics, University of Chicago, 515 CHSC, 920 E. 58th Street, Chicago, IL 60637, USA
**Correspondence to:* University of Chicago only; Tel: + 1 773 834 5239; Fax. + 1 773 834 0505; E-mail: vclark@genetics.bsd.uchicago.edu

Date received (in revised form): 16th January, 2004

Abstract

Chemokine signals and their cell-surface receptors are important modulators of HIV-1 disease and cancer. To aid future case/control association studies, aim to further characterise the haplotype structure of variation in chemokine and chemokine receptor genes. To perform haplotype analysis in a population-based association study, haplotypes must be determined by estimation, in the absence of family information or laboratory methods to establish phase. Here, test the accuracy of estimates of haplotype frequency and linkage disequilibrium by comparing estimated haplotypes generated with the expectation maximisation (EM) algorithm to haplotypes determined from Centre d'Etude Polymorphisme Humain (CEPH) pedigree data. To do this, they have characterised haplotypes comprising alleles at 11 biallelic loci in four chemokine receptor genes (*CCR3*, *CCR2*, *CCR5* and *CCRL2*), which span 150 kb on chromosome 3p21, and haplotyes of nine biallelic loci in six chemokine genes [*MCP-1(CCL2)*, *Eotaxin(CCL11)*, *RANTES(CCL5)*, *MPIF-1(CCL23)*, *PARC(CCL18)* and *MIP-1α(CCL3)*] on chromosome 17q11–12. Forty multi-generation CEPH families, totalling 489 individuals, were genotyped by the TaqMan 5′-nuclease assay. Phased haplotypes and haplotypes estimated from unphased genotypes were compared in 103 grandparents who were assumed to have mated at random.

For the 3p21 single nucleotide polymorphism (SNP) data, haplotypes determined by pedigree analysis and haplotypes generated by the EM algorithm were nearly identical. Linkage disequilibrium, measured by the D′ statistic, was nearly maximal across the 150 kb region, with complete disequilibrium maintained at the extremes between *CCR3*-Y17Y and *CCRL2*-I243V. D′-values calculated from estimated haplotypes on 3p21 had high concordance with pairwise comparisons between pedigree-phased chromosomes. Conversely, there was less agreement between analyses of haplotype frequencies and linkage disequilibrium using estimated haplotypes when compared with pedigree-phased haplotypes of SNPs on chromosome 17q11–12. These results suggest that, while estimations of haplotype frequency and linkage disequilibrium may be relatively simple in the 3p21 chemokine receptor cluster in population samples, the more complex environment on chromosome 17q11–12 will require a higher resolution haplotype analysis.

Keywords: *chemokine, SNP, haplotype estimation, pedigree analysis, linkage disequilibrium*

Introduction

Chemokines are small intercellular signalling molecules that recruit immune cells to sites of inflammation and infection. The two major subfamilies of chemokine proteins are defined as CC, with two adjacent cysteine residues, or as CXC, with an intervening non-conserved amino acid. Other chemokine family members have cysteine residues separated by more than one intervening amino acid (eg CX3C or Fractalkine),[1,2] or are characterised by having only one cysteine (eg XCL1 or Lymphotactin).[3,4] Chemokine receptors are defined by the subfamily of chemokine ligand that they bind. Both the chemokine and the chemokine receptor genes are generally clustered in four distinct chromosomal regions: CC on 17q11–21, CXC on 4q12–21, both CCR and CXCR on 3p21–24 and CXCR on 2q21–35.

Variation in chemokines, or their cell-surface receptors, influences an individual's susceptibility to HIV-1 infection and modulates progression to AIDS.[5–11] Chemokine signals are also important in the angiogenic[12–14] and metastatic[15,16] processes of cancer. Therefore, describing the genetic variation and haplotype structure of chemokine and chemokine receptor gene clusters is necessary for further disease association analyses of these candidate genes.

The focus of the present analysis is to describe the structure of multi-single nucleotide polymorphism (SNP) haplotypes in chemokine genes on chromosome 17q11–12 and chemokine receptor genes on chromosome 3p21 in Centre d'Etude Polymorphisme Humain (CEPH) pedigrees ($n = 489$). Secondary to this goal is to use the empirically phased haplotypes to determine the accuracy of estimated measures of haplotype frequencies and linkage disequilibrium using the subset of CEPH grandparents ($n = 103$).

Samples and methods

Study samples

SNP screening and validation were performed using two population panels: a 16-individual panel (four European-Americans, four African-Americans, four Chinese and four self-identified Hispanic-Americans) and an 88-individual panel (30 African-Americans, 34 European-Americans and 24 Hispanics). Forty multi-generation CEPH families, a total of 489 individuals, were genotyped for 23 SNPs scattered over two gene clusters: CC-chemokines on 17q11–12 and CC-chemokine receptors on 3p21 (see Table 1). Genotype data from a subsample of 103 unrelated grandparents were used for comparative haplotype analyses. The use of all anonymous DNA samples was either reviewed by the NIH Internal Review Board or determined 'exempt' from review.

Chemokine and chemokine receptor SNPs

Conditions for SNP detection in the CCR2 promoter. Four of the 23 SNPs included in the haplotype analysis (Table 1) have not previously been reported and were discovered by direct sequencing. Three kilobases of the *CCR2* promoter region were amplified using the Invitrogen Platinum Taq™ kit in a panel of 16 individuals (32 chromosomes), including four European-Americans, four African-Americans, four Chinese and four Hispanic-Americans (self-identified). For 100 μL polymerase chain reactions (PCRs), 200 nM deoxyribonucleotide triphosphates (dNTPs), 200 nM of each primer, 400 nM $MgSO_4$, 10 μL of 10 × Platinum Taq™ buffer and 1 μL Platinum Taq™ were mixed with approximately 100 ng of genomic DNA. Primer sequences for the 3 kb product were as follows: 5′-TCATCTGCTTCTTAATTGCCTTCAG-3′ (forward) and 5′-CAGGGTTTCTCTAACATCTCCTGGT-3′ (reverse). PCR was performed in a PE Biosystems 9700 ThermoCycler with long-range PCR conditions recommended for Platinum Taq™.

Sequencing was performed on a 3 kb segment at intervals of 400–500 kb with internal primers using the BigDye™ (Applied Biosystems) cycle sequencing kit with some modifications. Sequencing reactions were performed as follows: 15–30 ng of purified product was added to 10 μL reaction solution, which included 2 μL of BigDye™ mix, 1 μL of standard 5 × dilution buffer, 1.1 μL of 0.5 μM primer stock and double-distilled water (ddH_2O) for the remaining volume. Reactions were cycled in a PE Biosystems 9700 thermo cycler under the following conditions: 95°C for five minutes, and 30 cycles of 95°C for 30 seconds, 50°C for ten seconds and 60°C for four minutes. All individuals were sequenced for the entire 3 kb in both forward and reverse directions on an ABI 3700 capillary sequencer. Sequence trace files were analysed by the Phred/Phrap/Consed system,[17–20] and PolyPhred was used to detect putative SNPs.[21]

Eight SNPs (−5983 G/A, −5047 G/T, −4866 G/C, −4599 T/G, −4419 C/T, −4338 A/T, −3433 T/C and −3232 C/T) were confirmed by visual inspection of the CCR2 promoter sequence of the 16-individual screening panel. Five of these SNPs (−5983 G/A, −5047 G/T, −4866 G/C, −4599 T/G and −3433 T/C) were validated by direct sequencing in a larger sample set that comprised 88 individuals from three populations: 30 African-Americans, 34 European-Americans and 24 self-described Hispanics. The other three SNPs were not validated in the larger sample set, as they are in nearly complete linkage disequilibrium with at least one of the five SNPs chosen for further study. Four of the five validated SNPs listed in Table 1 (−5983 G/A, −5047 G/T, −4866 G/C and −3433 T/C) were successfully optimised for 5′-nuclease assays.

Conditions for screening putative SNPs. The remaining 19 SNPs listed in Table 1 were previously characterised in this laboratory by denaturing high performance liquid chromatography (dHPLC) or single-stranded conformation polymorphism (SSCP) analysis, or were taken from published works or public databases. Flanking primers were designed for a total of 22 polymorphisms from dbSNP[22] using Primer 3.0 from MIT, Cambridge, MA.[23] PCR was performed in 25 μL-scale reactions with the following components: 50 ng genomic DNA, 3 mM $MgCl_2$, 200 nM dNTPs, 200 nM of each primer, 1U TaqGold™ (Applied Biosytems) and 2.5 μL 10 × TaqGold™ Buffer. The cycling conditions (PE Biosystems 9700) for all reactions were as follows: a 95°C hold for ten minutes, then a touch-down cluster of 12 cycles (95°C for 30 seconds, 62–57°C (decreasing by 0.5°C every cycle) for one minute and 72°C for 1 minute), a standard cluster of 30 cycles (95°C for 30 seconds, 57°C for one minute and 72° for one minute) and a final 72°C hold for seven minutes. PCR products were purified using 10 U exonuclease 1 and 2 U shrimp alkaline phosphatase (SAP) enzymes under the protocol specified by the Washington University Sequencing Center.[24]

All purified reaction solutions were sequenced as follows: 15–30 ng of purified product was added to 10 μL reaction solution, which included 2 μL of BigDye™ mix, 1 μL of standard 5 × dilution buffer, 1.1 μL of 0.5 μM primer stock and ddH_2O for the remaining volume. Reactions were cycled in a PE Biosystems 9700 thermo cycler under the following

Table 1. Biallelic loci typed in CEPH pedigrees

Haplo-type position	Gene	NCBI locus link	Nucleotide/AA position	NCBI genbank number	NCBI contig	Contig position	NCBI dbSNP ss#	Allele 1	CEPH GP frequency
3p21									
1	CCR3	1232	Y17Y, A/G	NM_001837	NT_05827	3997337	4987053	A	0.907
	CCR2	1231	−5983 G/A	U95626	NT_05827	4083672		G	1
2	CCR2	1231	−5048 G/T	U95626	NT_05827	4084607	3918357	G	0.892
	CCR2	1231	−4866 C/G	U95626	NT_05827	4084789	3918370	C	1
3	CCR2	1231	−3433 T/C	U95626	NT_05827	4086222	3092964	T	0.802
4	CCR2	1231	V64I, C/T	NM_000647	NT_05827	4089845	1799864	C	0.898
5	CCR2	1231	N260N, A/G	NM_000647	NT_05827	4090435	1799865	A	0.696
6	CCR5	1234	208 C/A	NM_000579	NT_05827	4102477	2734648	G	0.631
7	CCR5	1234	303 C/T	NM_000579	NT_05827	4102572	1799987	G	0.524
8	CCR5	1234	676 T/C	NM_000579	NT_05827	4102945	1800023	A	0.631
9	CCR5	1234	L55Q, T/A	NM_000579	NT_05827	4105194	1799863	T	0.976
10	CCR5	1234	D32	NM_000579	NT_05827			NODEL	0.905
11	CCRL2	9034	I243V, C/T	NM_003965	NT_05827	4140934	3204850	C	0.902
	CCRL2	9034	1137 C/G	NM_003965	NT_05827	4141344		C	1
17q11–12									
1	MCP-1(CCL2)	6347	−362 C/G	M37719	NT_010799	7315787	2857656	C	0.718
2	EOTAXIN (CCL11)	6356	−1382 C/T	Z92709	NT_010799	7345226	4795895	C	0.777
3	RANTES (CCL5)	6352	−8147 A/G	NM_002985	NT_010799	8932972		A	0.903
4(1)	MPIF-1 (CCL23)	6368	M106V, G/A	U85767	NT_010799	9074064	1003645	A	0.832
5(2)	PARC (CCL18)	6362	−116 C/T	AB012113	NT_010799	9125397	2015086	C	0.662
6(3)	PARC (CCL18)	6362	81 G/A	AB012113	NT_010799	9125563	2015070	G	0.97
7(4)	PARC (CCL18)	6362	311 C/A	AB012113	NT_010799	9125793	2015052	A	0.922
8(5)	PARC (CCL18)	6362	6793 A/G	AB012113	NT_010799	9132275	14304	G	0.909
9(6)	MIP-1A (CCL3)	6348	−1541 T/C	M23178	NT_010799	9152727	1634497	A	0.705

conditions: 95°C for five minutes, and 30 cycles of 95°C for 30 seconds, 50°C for ten seconds, and 60°C for four minutes. Nine of the 22 primer pairs produced viable sequences and the SNPs were polymorphic in at least one of the 16-individual population panel. Those 'confirmed' SNPs were further characterised by either sequencing or genotyping in the larger sample of 88 individuals (data not shown).

SNP genotyping

All 23 SNPs were genotyped using the 5′-nuclease assay under a set of universal assay conditions. Dual-labelled TaqMan™ (Applied Biosystems) probes, standard, Turbo and Minor-Groove Binding (MGB) chemistries were designed using Primer Express™ (Applied Biosystems). Previous analysis of genotyping accuracy using the TaqMan method revealed 14 discordancies out of 1,165 duplicate genotype pairs, a 1.2 per cent error rate averaged over multiple TaqMan assays.[25] PCR conditions for genotyping (reaction components and cycling conditions), as described in Morin *et al.* (1999) and Clark *et al.* (2001), were used for all SNPs typed in this study.[25,26] PCR was performed in 96-well plates that included positive genotypic controls (for both homozygote states and the heterozygote state for each SNP) and reactions with no DNA as a negative control. All 5′-nuclease assay plates were read on the ABI 7700 Sequence Detector, and analysed using the 'dye components' feature of the SDS v1.6.3 or v1.7 software package (Applied Biosystems). Genotype determinations for each reaction were made manually by visual inspection of a scatter-plot of the data, with reference to the results of the genotype control samples. CEPH pedigree data for all 23 genotyping assays were checked for concordance with Mendelian inheritance using PEDCHECK.[27]

Haplotype analysis

Haplotype phase was determined using the CYRILLIC II pedigree drawing software (Cherwell Scientific) to establish the inheritance of multi-locus genotypes. The algorithm developed by Guo and Thompson (1992) was used to determine whether the distribution of whole haplotypes in the CEPH grandparent sample ($n = 103$) deviates from Hardy–Weinberg proportions.[28] Significance is determined by an exact test, with a cut-off of $p = 0.05$. Haplotype states and frequencies on both chromosomes 3p21 and 17q11–12 were estimated in sets of unphased genotype data by MLOCUS,[29,30] which uses the expectation-maximization (EM) algorithm,[31] a maximum-likelihood based method. A previously described three-step procedure to determine the most likely set of haplotypes to describe the genotype data was used here to analyse the haplotype states and frequencies for all datasets.[32] Haplotype blocks on 3p21 were assessed using HaploBlock-Finder,[33] which performs the four-gamete test (FGT) between each pairwise SNP to identify past recombination events.[34] The minimum-D′ method[35,36] (with minimum $D' = 0.80$) was also used to assess haplotype block structure in the 150 kb region of 3p21.

Validation of haplotype estimation

Haplotype frequencies are determined by direct counting of whole chromosomes in the grandparents after haplotypes are established by pedigree analysis. Haplotypes were estimated using MLOCUS with unphased genotype data from these same individuals. Comparisons of the two methods were performed with genotype data from two regions: the chemokine cluster (six genes) on chromosome 17q11–12 and the chemokine receptor cluster (four genes) on chromosome 3p21. For the 17q11–12 data, two analyses were performed: one included all nine SNPs typed in all six genes arrayed over 2 Mb, and the other included only six of these SNPs in the 77 kb 'core' region of three genes (*MPIF-1, PARC, MIP-1a*) on 17q11–12. The analysis of the 3p21 chemokine receptor genes included 14 SNPs arrayed over 150 kb.

The I_F and I_H algorithm performance indices suggested by Excoffier and Slatkin (1995) were used to quantitatively evaluate the estimation results in the CEPH grandparents.[37] The I_H index evaluates the performance of the algorithm to identify the actual haplotypes, and the I_F statistic examines how close the estimated frequencies are to the pedigree haplotype frequencies. I_H and I_F values were calculated using only those haplotypes above the threshold frequency (1/2n). A mean squared error (MSE) statistic was also used to compare the estimated haplotype frequencies to the pedigree-derived frequencies.[38] To determine whether omitting those grandparents who could not be phased from the analysis generates skewed pedigree-derived haplotype frequencies, MLOCUS haplotype estimations of the total sample ($n = 103$) were compared to the 'phased-only' sample using the above-described performance indices.

Estimating linkage disequilibrium in population data

D′ statistics were calculated with phased haplotypes derived from pedigree analysis with DnaSP (v3.53).[39] Linkage disequilibrium estimates generated by haplotypes determined by pedigree analysis in the CEPH grandparents were compared with those estimates calculated from MLOCUS reconstructed haplotypes in the same datasets. PAIRWISE was used to estimate linkage disequilibrium from the estimated haplotypes generated by MLOCUS.[30] PAIRWISE generates Lewontin's normalised D′ statistic[40] and the p-value determined from an exact test of association between all pairs of polymorphic loci in the dataset.

Results

Haplotype analysis of 3p21 SNPs

To determine the haplotype structure of SNPs in the 3p21 region, we typed 14 polymorphisms in the CEPH pedigrees.

Eleven of the 14 SNPs were polymorphic and none of these SNPs deviated from Hardy–Weinberg equilibrium (HWE) at the p = 0.05 significance level. Haplotype phase was established for every grandparent in the sample (n = 103). Haplotype frequencies were then determined by direct counting of whole haplotypes (Table 2). Nine haplotypes explained nearly all of the variation (98 per cent) in the CEPH grandparents. The remaining 2 per cent is composed of two haplotypes that occur only once.

The diplotypes, or multi-locus genotypes, were also counted in the CEPH grandparent sample. The diplotype combination of haplotypes 1 and 3 was the most frequent in the sample, at 13 per cent. In the CEPH grandparent sample, the 3p21 haplotypes were in HWE, as the randomisation test of the distribution of diplotypes yielded a non-significant p-value of 0.2708. When analysed individually, the 11 polymorphisms demonstrated no deviations from HWE in the CEPH grandparent sample.

Both haplotype block tests, the FGT and the minimal-D' method (set to the default of a minimum D' = 0.80), found a break between *CCR2*-N260N and *CCR5*-208. This indicates a past recombination event somewhere in the 20 kb between *CCR2* and *CCR5*. The pedigee haplotypes support this, as although there was no direct observation of a recombination event in the pedigree data, one haplotype (11112121211121) appeared to be a recombinant of haplotypes 4 (211111121211121) and 7 (11112121111111).

Haplotype analysis of 17q11-12 SNPs

To characterise the chemokine loci on chromosome 17, haplotype analysis was performed using all nine SNPs (over a 2 Mb region), as well as a subset of six SNPs arrayed over the 73 kb region, which includes *MPIF-1*, *PARC* and *MIP-1a*. Conclusive phase was established for only 87 individuals of the 103 in the CEPH grandparent sample for nine-SNP haplotypes. A total of 70 per cent of the variation of the total sample (n = 103) was explained by 14 haplotypes (of nine SNPs) (Table 3). The remaining portion included 11 doubleton haplotypes (found in two individuals), ten singletons (occured only once), as well as the 32 unphased chromosomes. When the analysis was reduced to six SNPs in the 73 kb region (Table 4), we were able to phase 96 grandparents by visual inspection of the pedigrees. Haplotype phase was not definitely assigned to seven of the 103 grandparents because two or more haplotype combinations could be inferred, given the diplotypes of their children or because of missing data. Eight six-SNP haplotypes explain 90 per cent of the variation in the CEPH grandparent sample (n = 103), and 41 per cent of the total number of chromosomes carry the most common haplotype (1 1 1 1 1 1) (Table 3). The remaining 10 per cent of the total number of chromosomes (2n = 206) is comprised of two doubletons, two singletons and the 14 unphased chromosomes.

Diplotypes were assigned to all individuals for which phase was established (n = 96) for the six SNPs in *MPIF-1*, *PARC* and *MIP-1a*. The most frequent diplotype combination

Table 2. Results from a comparison of pedigree-derived and estimated haplotype frequencies (n = 103). The total similarity index (I_F) and mean squared error (MSE) values are indicated at the bottom of the table. Haplotypes that are only present in MLOCUS estimates are denoted in italics. The haplotype number indicates the equivalent haplotype to those seven SNP haplotypes discussed in Clark *et al.* 2001.[25]

Haplotype number	Haplotype	GP count	Pedigree frequency	MLOCUS frequency	Similarity index	MSE
1	11111111111111	60	0.2913	0.2936	0.0024	0.00001
2	11111112221111	39	0.1893	0.1909	0.0015	0.00000
3	11112122221111	35	0.1699	0.1719	0.0020	0.00000
5	11211211111111	21	0.1019	0.1001	0.0019	0.00000
4	21111121211121	20	0.0971	0.0882	0.0089	0.00008
6	11111111111211	18	0.0874	0.0919	0.0045	0.00002
8	11111111112111	5	0.0243	0.0223	0.0020	0.00000
7	11112121111111	4	0.0194	0.0186	0.0008	0.00000
	11112121211121	2	0.0097	0.0098	0.0001	0.00000
	11212122221111	1	0.0049	0.0049	0.0000	0.00000
	11111121211111	1	0.0049	0.0049	0.0000	0.00000
	11211211211111	*0*	*0.0000*	*0.0022*	*0.0022*	*0.00000*
		206	**1.0000**	**0.9993**	**0.9869**	**0.00001**

Table 3. Comparison of pedigree-phased haplotypes for nine SNPs over 2 Mb of 17q11-12 in Centre d'Etude Polymorphisme Humain (CEPH) grandparents (n = 87) with MLOCUS estimates from unphased genotype data from these same individuals. The I_F (similarity index) and the mean squared error (MSE) for the two haplotype analyses are indicated at the bottom of the table. Those haplotypes present only in the MLOCUS analysis (less than 1 per cent frequency) are not included.

Haplotype	Count	Frequency	MLOCUS	Similarity index	MSE
111111111	36	0.2069	0.2374	0.0305	0.0009
111111122	22	0.1264	0.1332	0.0067	0.0000
121111111	19	0.1092	0.1450	0.0358	0.0013
211111111	18	0.1034	0.0631	0.0404	0.0016
211111122	9	0.0517	0.0605	0.0088	0.0001
111211111	8	0.0460	0.0245	0.0214	0.0005
111111121	6	0.0345	0.0341	0.0004	0.0000
121122111	5	0.0287	0.0145	0.0143	0.0002
121111121	4	0.0230	0.0198	0.0032	0.0000
121211122	3	0.0172	0.0000	0.0172	0.0003
211122111	3	0.0172	0.0315	0.0143	0.0002
112111122	3	0.0172	0.0168	0.0004	0.0000
112111111	3	0.0172	0.0196	0.0023	0.0000
121111122	3	0.0172	0.0209	0.0037	0.0000
121211111	2	0.0115	0.0051	0.0064	0.0000
212111122	2	0.0115	0.0000	0.0115	0.0001
212111111	2	0.0115	0.0345	0.0230	0.0005
211221211	2	0.0115	0.0165	0.0050	0.0000
211111121	2	0.0115	0.0088	0.0027	0.0000
211211111	2	0.0115	0.0218	0.0103	0.0001
111121211	2	0.0115	0.0000	0.0115	0.0001
122211111	2	0.0115	0.0113	0.0002	0.0000
111122111	2	0.0115	0.0000	0.0115	0.0001
122111122	2	0.0115	0.0000	0.0115	0.0001
111211122	2	0.0115	0.0245	0.0130	0.0002
111111112	1	0.0057	0.0000	0.0057	0.0000
112211111	1	0.0057	0.0000	0.0057	0.0000
121222111	1	0.0057	0.0000	0.0057	0.0000
211211122	1	0.0057	0.0000	0.0057	0.0000
111221211	1	0.0057	0.0000	0.0057	0.0000
121111112	1	0.0057	0.0066	0.0009	0.0000

(*continued*)

Table 3. *Continued.*

Haplotype	Count	Frequency	MLOCUS	Similarity index	MSE
111222111	1	0.0057	0.0136	0.0078	0.0001
122111111	1	0.0057	0.0000	0.0057	0.0000
211122122	1	0.0057	0.0000	0.0057	0.0000
111121212	1	0.0057	0.0000	0.0057	0.0000
	174	**1.0000**	**0.9636**	**0.8196**	**0.0002**

included haplotypes 1 and 2, at 28 per cent in the CEPH grandparents. There was no significant deviation from Hardy–Weinberg proportions for the six-SNP multi-site genotypes, with a randomisation p-value of 0.1102. When analysing the SNPs individually for HWE, one SNP — *PARC* (−116) — showed a significant deviation using a χ^2 test, at $p = 0.012$, which did not survive a Bonferroni multiple-test correction.

Validation of the EM algorithm on 3p21 and 17q11–12

To validate the accuracy of the EM algorithm, we compared the pedigree-derived haplotypes to those estimated haplotypes generated by MLOCUS. The 3p21 haplotype distributions were nearly identical to the estimated frequencies (Table 2). The similarity (I_F) and identity (I_H) indices were calculated for haplotypes in the CEPH grandparent sample ($n = 103$) for 14 SNPs. For the 14 SNP haplotypes in 3p21, as indicated in the Table 2, the similarity index (I_F) was 0.9869. An I_F of 1.0 would indicate perfect concordance between the haplotype frequencies generated by the two methods. The identity index (I_H) for these data was exactly 1.0, as all haplotypes derived by pedigree analysis were present in the MLOCUS results. One estimated haplotype was dropped from the analysis, as it was below the frequency threshold of ($1/2n = 0.004854$), as suggested by Excoffier and Slatkin (1995).[37] The MSE incorporates the overall difference in frequencies between actual (pedigree-derived) and estimated frequencies for all *H* haplotypes. The MSE for the 3p21 haplotypes was small (0.00001), which, again, indicates that the two frequency distributions are nearly identical.

As mentioned previously, phase could not be determined for the nine SNPs typed on chromosome 17q11–12 for all grandparents. Haplotype frequencies were determined, both by whole chromosome counting and by estimation, with data from 87 out of 103 individuals. The similarity index (I_F) for the distribution of frequencies for the 43 haplotypes (nine SNPs) in this region is 0.8196, as indicated in Table 3. The haplotype estimation yielded 24 haplotypes with frequencies over the threshold value ($1/(2n) = 0.0057$), and missed 13 haplotypes that were present in the pedigree data. The I_H statistic for these data is 0.7457. The MSE for the nine-SNP haplotypes is 0.0002, as indicated in Table 3. The EM algorithm also generated seven low frequency haplotypes (less than 1 per cent, not shown) that were not observed in the pedigree analysis. Constraining the MLOCUS analysis by removing these haplotypes did not significantly improve the MLE. This constrained analysis also resulted in the generation of other spurious low-frequency haplotypes, indicating that the EM algorithm could not effectively resolve haplotype phase for some individuals in the nine-SNP dataset.

Not surprisingly, paring the analysis down to the six SNPs in the 77 kb region that contains *MPIF-1*, *PARC* and *MIP-1α* yields more accurate haplotype estimates. Ninety-six grandparents were included in this analysis, as phase could not be determined for seven of the 103 individuals in the total sample. As indicated in Table 4, the I_F statistic increased to 0.9491, and the I_H of 0.9167 is closer to perfect identity (1.0). The MSE is also closer to zero, at 0.0001.

Comparisons of MLOCUS haplotype estimates for 17q11–12

Omitting the unphased chromosomes from the pedigree haplotype frequency calculation of the 17q11–12 SNPs is a potential source of bias, as those individuals for whom complete resolution is not possible may have a higher per site heterozygosity than randomly sampled individuals. Additionally, those ‘unphasable’ individuals may carry haplotypes that are not present in the phased portion of the sample. To test if using only the phased individuals generates skewed ‘pedigree-derived’ 17q11–12 haplotype frequencies, MLOCUS haplotype frequency estimates were generated from both the total dataset of unphased genotypes ($n = 103$, data not shown) and those genotypes only from the phased individuals — $n = 87$, for the nine-SNP haplotypes (Table 3), and $n = 96$ for the six-SNP haplotypes (Table 4). Comparisons of nine-SNP MLOCUS haplotypes (above 1 per cent frequency) from the whole sample ($n = 103$) and the phased sample ($n = 87$) yielded an I_H of 0.9729, an I_F of 0.9313 and an MSE of 0.00007. The same comparison performed on the six-SNP haplotypes yielded an I_H of 1, an I_F of 0.9838 and an MSE of 0.00002. One nine-SNP haplotype present in the total sample (at a frequency of 0.015) was missed in the ‘phased-only’ sample, while in the six-SNP analysis, both sets of genotypes

Table 4. Comparison of MLOCUS estimated to pedigree-phased haplotypes (n = 96) for six SNPs in 79 kb 'core' region of 17q11-12. Those haplotypes that are only present in the MLOCUS estimation results are denoted in italics.

No.	HAPLOTYPE	GP count	Frequency	MLOCUS	Similarity index	MSE
1	111111	85	0.4427	0.4519	0.0092	0.00008
2	111122	46	0.2396	0.2632	0.0236	0.00056
3	211111	20	0.1042	0.0934	0.0108	0.00012
4	111121	12	0.0625	0.0610	0.0015	0.00000
5	122111	10	0.0521	0.0503	0.0018	0.00000
6	211122	6	0.0313	0.0148	0.0164	0.00027
7	221211	4	0.0208	0.0254	0.0045	0.00002
8	111112	3	0.0156	0.0063	0.0093	0.00009
9	121211	2	0.0104	0.0046	0.0058	0.00003
10	222111	2	0.0104	0.0174	0.0070	0.00005
11	121212	1	0.0052	0.0065	0.0013	0.00000
12	122122	1	0.0052	0.0000	0.0052	0.00003
13	*211112*	*0*	*0.0000*	*0.0052*	*0.0052*	*0.00003*
		192	**1.0000**	**1.0000**	**0.9491**	**0.00010**

generated identical haplotypes. The potential bias of removing the unphased grandparents from the haplotype analyses appears to be slight, as the index values indicate that the haplotype frequencies generated by the two datasets (the complete sample and the 'phased-only' sample) are very similar, particularly for the six-SNP haplotypes.

Comparisons of methods to estimate linkage disequilibrium

Both phased haplotypes and unphased genotype data from the CEPH grandparents (n = 103) were used to estimate the extent of pairwise linkage disequilibrium (described by D') between SNPs in the chemokine receptor region on chromosome 3p21 and the chemokine cluster on chromosome 17q11–12. The D' statistic (above the diagonal) and the measure of statistical significance (p-value) (below the diagonal) are presented for pairwise comparisons of the 11 polymorphic sites in 3p21 in Table 5. Negative values indicate that there is disequilibrium between opposite alleles at the two SNPs (ie allele 1 at the first SNP and allele 2 at the second SNP, where the common allele is allele 1).

The D'-values generated from analyses of the 3p21 polymorphisms by the DnaSP and PAIRWISE programs were, for the most part, very similar. The three differences, noted in bold, are slight. As discussed previously, the haplotypes generated by the EM algorithm were essentially identical when compared with those discerned by pedigree analysis for the variants in this region. The analysis of both the haplotypes and the unphased genotype data indicated that linkage disequilibrium in this 150 kb region of 3p21 is high in the CEPH grandparents. There is intact linkage disequilibrium ($D' = 1$) between two SNPs at the extremes of the region (*CCR3*-Y17Y and *CCRL2*-I243V), preserved primarily on haplotype 4 (2 1 1 1 1 1 1 2 1 2 1 1 1 2 1). The relative loss of linkage disequilibrium in the centre of the region, between *CCR2*-N260N and two SNPs in the CCR5 promoter, 208 ($D' = 0.326$) and CCR5-676 ($D' = 0.326$), was detected by haplotype block analysis, indicating past recombination between these two genes.

It is not surprising that the DnaSP analysis of haplotypes on 17q11–12 indicated no evidence of long-range linkage disequilibrium between variants at the extremes of the 2 MB region. There is significant linkage disequilibrium between the SNPs typed in *MCP-1* and nearby *Eotaxin* ($D' = -1$) at the centromeric end of the region. Likewise, there is some significant allelic association between SNPs in *MIP-1α*, *PARC* and *MPIF-1*, which are within 77 kb of each other. The relative lack of association between more distal SNPs seems to have hampered the ability of the PAIRWISE analysis of unphased genotype data to accurately detect the extent of linkage disequilibrium, when compared with the DnaSP analysis of whole haplotypes. This lack of sensitivity is especially evident in the analyses of all nine SNPs, as the

Table 5. Estimated D′ values generated by two methods for all polymorphic loci in the 3p21 chemokine receptor gene region in the CEPH sample. Numbers above the diagonal for each table indicate the pairwise D′ value for pairs of SNPs in the CEPH grandparent sample (n = 103). p values for each test are indicated below the diagonal. Values in the upper table (A) indicate D′ values calculated in DnaSP from the pedigree-derived haplotypes. Values in the lower table (B) indicate D′ values calculated using the PAIRWISE program from the MLOCUS haplotype frequency estimates. Those values in boxed cells indicate non-significant results. D′ estimates in the lower table denoted in italics indicate differences from the values generated in DnaSP for that particular comparison of loci.

	Y17Y	−5048	−3433	V64I	N260N	208	303	676	L55Q	D32	I243V
A. DnaSP											
CCR3(Y17Y)		−1	−1	−1	1	−1	1	−1	−1	−1	1
CCR2(−5048)	0.138		−0.777	1	−0.851	−0.875	−0.904	−0.875	−1	−1	−1
CCR2(−3433)	0.016	0.053		−1	1	0.775	0.818	0.775	−1	−1	−0.573
CCR2(V64I)	0.142	<0.001	0.009		−1	−1	−1	−1	−1	−1	−1
CCR2(N260N)	<0.001	0.003	<0.001	<0.001		0.326	0.879	0.326	−1	−1	1
CCR5(208)	<0.001	0.001	<0.001	<0.001	<0.001		1	1	−1	−1	−1
CCR5(303)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001		1	−1	−1	1
CCR5(676)	<0.001	0.001	<0.001	<0.001	<0.001	<0.001	<0.001		−1	−1	−1
CCR5(L55Q)	1	1	0.586	1	0.189	0.161	0.061	0.161		−1	−1
CCR5(D32)	0.227	0.228	0.017	0.226	0.001	<0.001	<0.001	<0.001	1		−1
CCRL2(I243V)	<0.001	0.086	0.176	0.139	<0.001	<0.001	<0.001	<0.001	1	0.143	
B. PAIRWISE											
CCR3(Y17Y)		1	1	1	1	−1	1	−1	−1	−1	1
CCR2(−5048)	0.123		*−0.777*	1	−0.847	−0.876	−0.903	−0.876	−0.163	−1	−1
CCR2(−3433)	0.024	0.05		−1	1	0.781	0.828	0.781	−1	−1	−0.513
CCR2(V64I)	0.133	<0.001	0.014		−1	−1	−1	−1	−0.123	−1	−1
CCR2(N260N)	<0.001	0.006	<0.001	0.002		0.355	0.882	0.355	−1	−1	1
CCR5(208)	0.001	0.001	<0.001	<0.001	<0.001		1	1	−1	−1	−1
CCR5(303)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001		1	*−1*	−1	1
CCR5(676)	0.001	0.001	<0.001	<0.001	<0.001	<0.001	<0.001		−1	−1	−1
CCR5(L55Q)	0.481	0.898	0.25	0.925	0.14	0.084	0.033	0.084		−1	−1
CCR5(D32)	0.157	0.115	0.021	0.125	0.003	0.001	<0.001	0.001	0.471		−1
CCRL2(I243V)	<0.001	0.103	0.219	0.112	<0.001	<0.001	<0.001	<0.001	0.455	0.134	

Table 6. Estimated D′ values generated by two methods for all nine SNPs in the 2 Mb chemokine gene region on chromosome 17q11-12 in CEPH grandparents (n = 87). Numbers above the diagonal for each table indicate the pairwise D′ value for pairs of SNPs in the CEPH grandparent sample. p values for each test are indicated below the diagonal. Values in the upper table (A) indicate D′ values calculated in DnaSP from the pedigree-derived haplotypes. Values in the lower table (B) indicate D′-values calculated using the PAIRWISE program from the MLOCUS haplotype frequency estimates. Those values in boxed cells indicate non-significant results. D′ estimates in the lower table denoted in italics indicate differences from the values generated in DnaSP for that particular comparison of loci.

	MCP-1	Eotaxin	RANTES	MPIF-1 (M106V)	PARC(−116)	PARC(81)	PARC(311)	PARC(6793)	MIP-1a
A. DnaSP									
MCP-1		−1	−0.025	−0.171	0.098	0.087	0.121	0.099	0.076
Eotaxin	<0.001		0.07	0.051	0.098	0.29	−1	−0.228	−0.375
RANTES	1	0.766		0.038	−1	−1	−1	−0.306	0.239
MPIF-1(M106V)	0.801	0.619	0.715		0.14	0.012	0.416	−0.174	−0.081
PARC(−116)	0.57	0.57	0.222	0.157		1	1	−0.379	−0.309
PARC(81)	0.736	0.085	0.369	1	<0.001		−1	−0.773	−0.747
PARC(311)	0.632	0.338	1	0.039	<0.001	1		0.243	0.281
PARC(6793)	0.351	0.264	0.426	0.649	0.305	0.038	0.409		0.829
MIP-1a	0.443	0.083	0.163	0.82	0.435	0.069	0.37	<0.001	
B. PAIRWISE									
MCP-1		−0.940	0.192	−0.060	*0.379*	0.297	*0.557*	0.110	0.072
Eotaxin	<0.001		0.033	0.244	0.102	0.199	−0.324	*−0.359*	*−0.378*
RANTES	0.146	0.803		0.116	−0.052	−0.031	−0.077	0.328	0.004
MPIF-1(M106V)	0.591	0.474	0.845		*0.288*	0.147	0.593	−0.154	−0.061
PARC(−116)	0.002	0.408	0.521	<0.001		1	1	−0.371	−0.077
PARC(81)	0.052	0.193	0.721	0.157	<0.001		−1	*−0.554*	−0.474
PARC(311)	0.015	0.643	0.560	<0.001	<0.001	0.478		0.014	0.323
PARC(6793)	0.546	0.049	0.303	0.567	0.214	0.132	0.962		0.935
MIP-1a	0.728	0.067	0.990	0.840	0.818	0.254	0.211	<0.001	

Table 7. Estimated D′ values generated by two methods for six SNPs in the 79 kb ‘core’ region of three chemokine genes on chromosome 17q11–12 in CEPH. Numbers above the diagonal for each table indicate the pairwise D′ value for pairs of SNPs in the CEPH grandparent sample (n = 96). p values for each test are indicated below the diagonal. Values in the upper table (A) indicate D′ values calculated in DnaSP from the pedigree-derived haplotypes. Values in the lower table (B) indicate D′ values calculated using the PAIRWISE program from the MLOCUS haplotype frequency estimates. Those values in boxed cells indicate non-significant results. D′ estimates in the lower table denoted in italics indicate differences from the values generated in DnaSP for that particular comparison of loci.

	MPIF-1(M106V)	PARC(−116)	PARC(81)	PARC(311)	PARC(6793)	MIP-1a(−1541)
A. DnaSP						
MPIF-1(M106V)		$\boxed{0.174}$	$\boxed{-0.105}$	0.597	$\boxed{-0.157}$	$\boxed{0.023}$
PARC(−116)	$\boxed{0.105}$		1	1	$\boxed{-0.268}$	$\boxed{-0.172}$
PARC(81)	$\boxed{1}$	<0.001		$\boxed{-1}$	−0.786	$\boxed{-0.758}$
PARC(311)	0.009	<0.001	$\boxed{1}$		$\boxed{0.48}$	$\boxed{0.511}$
PARC(6793)	$\boxed{0.552}$	$\boxed{0.454}$	0.034	$\boxed{0.19}$		0.898
MIP-1a	$\boxed{0.839}$	$\boxed{0.796}$	$\boxed{0.065}$	$\boxed{0.082}$	<0.001	
B. PAIRWISE						
MPIF-1(M106V)		*0.302*	$\boxed{-0.120}$	0.640	*−0.720*	*0.566*
PARC(−116)	0.001		1	1	−1	$\boxed{-0.789}$
PARC(81)	$\boxed{0.298}$	<0.001		$\boxed{-1}$	−1	*−1*
PARC(311)	<0.001	<0.001	$\boxed{0.468}$		$\boxed{0.559}$	$\boxed{0.398}$
PARC(6793)	0.002	0.001	0.007	$\boxed{0.054}$		0.908
MIP-1a	0.029	$\boxed{0.016}$	0.015	$\boxed{0.486}$	<0.001	

multitude of haplotypes (including spurious haplotypes generated by the EM estimation) created false-positive associations between distal variants (such as between *MCP-1* and SNPs in *PARC*) (Tables 6 and 7).

Discussion

Given the potential accuracy of low-cost statistical methods, and the current high cost of molecular haplotyping and pedigree analysis, statistical estimation to determine haplotypes may be a cost-effective strategy for many gene regions. As a minimum, statistical estimation can be used to determine the overall need for molecular haplotyping and to specify where in the dataset molecular haplotyping would provide the most benefit.[41–43] Independent assessments of the effectiveness of the EM algorithm have been discussed at length.[38,44,45] Xu *et al.* (2002)[46] discuss a comparison of three computational algorithms for estimating haplotype frequencies: the Clark (1990) rule-based method,[47] the EM algorithm and the Stephens *et al.* (2001) Bayesian PHASE method.[48] Using previously described criteria,[37] Xu *et al.* found that all three methods performed better for regions with a high degree of linkage disequilibrium, such as in the *NAT2* gene, than for regions where linkage disequilibrium is not maintained (chromosome 8p22) when compared with haplotypes determined by molecular methods.[46]

The purpose of the evaluation presented here is to establish the accuracy of statistical estimation in these chemokine and chemokine receptor gene clusters. Estimated haplotypes from unphased genotypes were compared with haplotypes derived empirically from pedigree analysis in the CEPH grandparent sample ($n = 103$). How the EM algorithm responds to irregular linkage disequilibrium, sample size, different levels of polymorphism and deviations from HWE is critical for the effectiveness of haplotype estimation.[38] These conditions will be affected by the genomic environment of the region of interest, the history of the population from which the samples were selected and the quality of the genotype data. While these validation results cannot control for all these variables, an attempt was made to explore how the EM algorithm responds to the conditions of the gene clusters studied on chromosomes 3p21 and 17q11–12 in a European-derived sample set.

A greater degree of linkage disequilibrium between SNPs, and therefore fewer haplotypes, increases the accuracy of the EM algorithm and aids subsequent estimates of measures of linkage disequilibrium (such as D'). This is evident from the results of estimations of haplotype frequency and linkage disequilibrium in the 150 kb region on 3p21. Relatively few haplotypes explain the variation between these SNPs, at least in the CEPH grandparent sample. Indeed, there is intact linkage disequilibrium at the extremes of this region, as *CCR3*-Y17Y and *CCRL2*-I243V have a pairwise D'-value of 1. The haplotype block analysis also indicates a fairly simple structure, as both tests applied here found only two blocks, with what appeared to be a past recombination event between *CCR2* and *CCR5*.

The degree of linkage disequilibrium between SNPs is one of the most important factors in the ability of the EM algorithm to properly detect haplotypes in population samples.[38,44] The analysis presented here shows that the EM algorithm accurately describes the haplotype structure and patterns of pairwise linkage disequilibrium on chromosome 3p21 (a region of higher linkage disequilibrium). As for chromosome 17, it is important to note that, because of the relatively few SNPs assessed (a total of nine), this analysis is a low resolution evaluation of haplotypes and linkage disequilibrium across a large region (2 Mb). While including only the 'core' region of 17q11–12 yields more accurate estimates of haplotype frequencies and linkage disequilibrium, these analyses still include a relatively sparse sampling of SNPs (six in 77 kb). The results of the pedigree analysis indicate that, while haplotype estimations in the chemokine receptor cluster on 3p21 may be fairly straightforward, special care must be taken for any haplotype inference in the chemokine genes on chromosome 17. More SNP genotype data, especially in the chromosome 17 chemokine genes, will no doubt aid in further characterisation of variation and linkage disequilibrium in these gene regions, as well as improve the accuracy of future haplotype analyses.

Acknowledgments

The authors would like to thank Bert Gold, Jim Lautenberger, Raymond Peterson and George Nelson for helpful discussion of haplotype analysis. Bill Modi, Noah Metheny and Julie Bergeron provided technical assistance with some TaqMan assays, and the LGD Cell Repository aided with DNA extraction. Jeff Long provided software and Carrie Pfaff performed necessary program modifications for this analysis. We also wish to thank the two anonymous reviewers for helpful comments.

References

1. Bazan, J.F, Bacon, K.B., Hardiman, G. *et al.* (1997), 'A new class of membrane-bound chemokine with a CX3C motif', *Nature* Vol. 385, pp. 640–644.
2. Pan, Y., Lloyd, C., Zhou, W. *et al.* (1997), 'Neurotactin, a membrane-anchored chemokine upregulated in brain inflammation [published erratum appears in Nature (1997), Vol. 389, p.100]', *Nature* Vol. 387, pp. 611–617.
3. Yoshida, T., Imai, T., Kakizaki, H. *et al.* (1995), 'Molecular cloning of a novel C or gamma type chemokine, SCM-1', *FEBS Lett.* Vol. 360, pp. 155–159.
4. Yoshida, T., Imai, T., Kakizaki, H. *et al.* (1998), 'Identification of single C motif-1/lymphotactin receptor XCR1', *J. Biol. Chem.* Vol. 273, pp. 16551–16554.
5. Cocchi, F., DeVico, A.L., Garzino-Demo, A. *et al.* (1995), 'Identification of RANTES, MIP-1 alpha, and MIP-1 beta as the major HIV-suppressive factors produced by CD8 + T cells', *Science* Vol. 270, pp. 1811–1815.
6. Dean, M., Carrington, M., Winkler, C. *et al.* (1996), 'Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study,

Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study', *Science* Vol. 273, pp. 1856–1862.

7. Smith, M.W., Carrington, M., Winkler, C. *et al.* (1997), 'CCR2 chemokine receptor and AIDS progression', *Nat. Med.* Vol. 3, pp. 1052–1053.
8. Kostrikis, L.S., Huang, Y., Moore, J.P. *et al.* (1998), 'A chemokine receptor CCR2 allele delays HIV-1 disease progression and is associated with a CCR5 promoter mutation', *Nature Med.* Vol. 4, pp. 350–353.
9. Martin, M.P., Dean, M., Smith, H.W., *et al.* (1998), 'Genetic acceleration of AIDS progression by a promoter variant of CCR5', *Science* Vol. 282, pp. 1907–1911.
10. Winkler, C., Modi, W., Smith, H.W. *et al.* (1998), 'Genetic restriction of AIDS pathogenesis by an SDF-1 chemokine gene variant. ALIVE Study, Hemophilia Growth and Development Study (HGDS), Multicenter AIDS Cohort Study (MACS), Multicenter Hemophilia Cohort Study (MHCS), San Francisco City Cohort (SFCC)', *Science* Vol. 279, pp. 389–393.
11. An, P., Martin, M.P., Nelson, G.W. *et al.* (2000), 'Influence of CCR5 promoter haplotypes on AIDS progression in African-Americans', *Aids* Vol. 14, pp. 2117–2122.
12. Strieter, R.M., Polverini, P.J., Arenberg, D.A. *et al.* (1995), 'Role of C-X-C chemokines as regulators of angiogenesis in lung cancer', *J. Leukoc. Biol.* Vol. 57, pp. 752–762.
13. Arenberg, D.A., Kunkel, S.L., Polverini, P.J. *et al.* (1996), 'Interferon-gamma-inducible protein 10 (IP-10) is an angiostatic factor that inhibits human non-small cell lung cancer (NSCLC) tumorigenesis and spontaneous metastases', *J. Exp. Med.* Vol. 184, pp. 981–999.
14. Moore, B.B., Arenberg, D.A., Addison, C.L. *et al.* (1998), 'Tumor angiogenesis is regulated by CXC chemokines', *J. Lab. Clin. Med.* Vol. 132, pp. 97–103.
15. Wang, J.M., Chertov, O., Proost, P. *et al.* (1998), 'Purification and identification of chemokines potentially involved in kidney-specific metastases by a murine lymphoma variant: Induction of migration and NFKB activation', *Inlt. J. Cancer* Vol. 75, pp. 900–907.
16. Muller, A., Homey, B., Soto, H. *et al.* (2001), 'Involvement of chemokine receptors in breast cancer metastasis', *Nature* Vol. 410, pp. 50–56.
17. Gordon, D., Abajian, C., Green, P. *et al.* (1998), 'Consed: A graphical tool for sequence finishing', *Genome Res.* Vol. 8, pp. 195–202.
18. Kwok, P.Y., Carlson, C., Yager, J.D. *et al.* (1994), 'Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products', *Genomics* Vol. 23, pp. 138–144.
19. Ewing, B. and Green, P. (1998), 'Base-calling of automated sequencer traces using phred. II. Error probabilities', *Genome Res.* Vol. 8, pp. 186–194.
20. Ewing, B., Hillier, L., Wendl, H.C. *et al.* (1998), 'Base-calling of automated sequencer traces using phred. II. Accuracy assessment', *Genome Res.* Vol. 8, pp. 175–185.
21. Nickerson, D.A., Tobe, V.O., Taylor, S.L., *et al.* (1997), 'PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing', *Nucleic Acids Res.* Vol. 25, pp. 2745–2751.
22. NTH (2001–2003), dbSNP, http://www.ncbi.nlm.nih.gov/SNP/, National Center for Biotechnology Information, National Institutes of Health.
23. Whitehead Institute (2001–2003), Primer 3.0, http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi, Whitehead Institute.
24. Genome Sequencing Centre (2001), Exo-SAP Protocol, http://genome.wustl.edu/tools/protocols/, Genome Sequencing Center, Washington University.
25. Clark, V.J., Metheny, N., Dean, M. and Peterson, R. (2001), 'Statistical estimation and pedigree analysis of CCR2-CCR5 haplotypes', *Hum. Genet.* Vol. 108, pp. 484–493.
26. Morin, P.A., Saiz, R. and Monjazeb, A. (1999), 'High-throughput single nucleotide polymorphism genotyping by fluorescent 5′ exonuclease assay', *Biotechniques* Vol. 544, pp. 538–540, 542, 544.
27. O'Connell, J.R. and Weeks, D.E. (1998), 'PEDCHECK: A program for identification of genotype incompatibilities in linkage analysis', *Am. J. Hum. Genet.* Vol. 63, pp. 259–266.
28. Guo, S.W. and Thompson, E.A. (1992), 'Performing the exact test of Hardy–Weinberg proportion for multiple alleles', *Biometrics* Vol. 48, pp. 361–372.
29. Long, J.C., Williams, R.C. and Urbanek, M. (1995), 'An E-M algorithm and testing strategy for multiple-locus haplotypes', *Am. J. Hum. Genet.* Vol. 56, pp. 799–810.
30. Long, J.C. (1999), 'Multiple locus haplotype analysis (MLOCUS, OBSHAP, PAIRWISE), software and documentation distributed by the author', Bethesda, MD, Section on Population Genetics and Linkage, Laboratory of Neurogenetics, NIAAA, National Institutes of Health.
31. Dempster, A.P. (1977), 'Maximum-likelihood from incomplete data via the EM algorithm', *J.R. Stat. Soc. B* Vol. 39, pp. 1–38.
32. Peterson, R.J., Goldman, D., *et al.* (1999), 'Effects of worldwide population subdivision on ALDH2 linkage disequilibrium', *Genome Res.* Vol. 9, pp. 844–852.
33. Zhang, K. and Jin, L. (2003), 'HaploBlockFinder: Haplotype block analyses', *Bioinformatics* Vol. 19, pp. 1300–1301.
34. Wang, N., Akey, J.M., Zhang, K. *et al.* (2002), 'Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation', *Am. J. Hum. Genet.* Vol. 71, pp. 1227–1234.
35. Daly, M.J., Rioux, J.D., Schaffner, S.F. *et al.* (2001), 'High-resolution haplotype structure in the human genome', *Nat. Genet.* Vol. 29, pp. 229–232.
36. Gabriel, S.B., Schaffner, S.F., Nguyen, H. *et al.* (2002), 'The structure of haplotype blocks in the human genome', *Science* Vol. 296, pp. 2225–2229.
37. Excoffier, L. and Slatkin, M. (1995), 'Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population', *Mol. Biol. Evol.* Vol. 12, pp. 921–927.
38. Fallin, D. and Schork, N.J. (2000), 'Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data', *Am. J. Hum. Genet.* Vol. 67, pp. 947–959.
39. Rozas, J. and Rozas, R. (1999), 'DnaSP version 3: An integrated program for molecular population genetics and molecular evolution analysis', *Bioinformatics* Vol. 15, pp. 174–175.
40. Lewontin, R.C. (1964), 'The interaction of selection and linkage. I. General considerations: Heterotic models', *Genetics* Vol. 49, pp. 49–67.
41. Michalatos-Beloin, S., Tishkoff, S.A., Bentley, K.L. *et al.* (1996), 'Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR', *Nucleic Acids Res.* Vol. 24, pp. 4841–4843.
42. Tishkoff, S.A., Dietzsch, E., Speed, W. *et al.* (1996), 'Global patterns of linkage disequilibrium at the CD4 locus and modern human origins', *Science* Vol. 271, pp. 1380–1387.
43. Clark, A.G., Weiss, K.M., Nickerson, D.A. *et al.* (1998), 'Haplotype structure and population genetic inferences from nucleotide sequence variation in human lipoprotein lipase', *Am. J. Hum. Genet.* Vol. 63, pp. 595–612.
44. Tishkoff, S.A., Pakstis, A.J., Ruano, G. and Kidd, K.K. (2000), 'The accuracy of statistical methods for estimation of haplotype frequencies: An example from the CD4 locus', *Am. J. Hum. Genet.* Vol. 67, pp. 518–522.
45. McKeigue, P.M. (2000), 'Efficiency of estimation of haplotype frequencies: Use of marker phenotypes of unrelated individuals versus counting of phase-known gametes', *Am. J. Hum. Genet.* Vol. 67, pp. 1626–1627.
46. Xu, W., Tse, H.F., Chan, F.H., *et al.* (2002), 'New Bayesian discriminator for detection of atrial tachyarrhythmias', *Circulation* Vol. 105, pp. 1472–1479.
47. Clark, A.G. (1990), 'Inference of haplotypes from PCR-amplified samples of diploid populations', *Mol. Biol. Evol.* Vol. 7, pp. 111–122.
48. Stephens, M., Smith, N.J., Donnelly, P. *et al.* (2001), 'A new statistical method for haplotype reconstruction from population data', *Am. J. Hum. Genet.* Vol. 68, pp. 978–989.