Population Genetics and Comparative Genetics of *CLDN1*, a Gene Involved in Hepatitis C Virus Entry

Vincent Bekker^a Thomas R. O'Brien^b Stephen Chanock^{a–c}

^aSection on Genomic Variation, Pediatric Oncology Branch, Center for Cancer Research, and ^bDivision of Cancer Epidemiology and Genetics, NCI, NIH, Bethesda, Md., and ^cCore Genotyping Facility, NCI, NIH, Gaithersburg, Md., USA

Key Words

Hepatitis C virus • Population genetics • Evolution • Selection • Conservation • Tight junctions • Viral receptor

Abstract

The claudin-1 gene (*CLDN1*) is a member of a family of genes that encodes proteins found in tight junctions and it has recently been implicated as one of several receptors for late stage binding of hepatitis C virus (HCV). Exploration of the population genetics of this gene could be informative, especially in the investigation of a possible genetic contribution to HCV infection. Comparison to a highly similar gene, claudin-7 (*CLDN7*) could provide insight into the recent molecular evolution of *CLDN1*. Mean interspecies conservation score was 0.11 (SD 0.28) for *CLDN1* and 0.31 (SD 0.43) for *CLDN7*. Re-sequence analysis was performed across all exons and evolutionarily conserved regions in *CLDN1* (13 kb in total) and *CLDN7* (2 kb in total) in 204 chromosomes drawn from the SNP500Cancer resource of four self-described ethnic groups in the US. For *CLDN1*, 133 SNPs were identified as well as 8 indels and an AC repeat length polymorphism. For *CLDN7*, 5 SNPs were identified. Assessment of nucleotide diversity (including F_{st} , θ and π statistics) did not show evidence for recent positive or negative selection in either gene. The pat-

tern of linkage disequilibrium was determined for each group and there is substantial difference for common SNPs (>5%) between populations as well as genes, further supporting the absence of signatures of recent selection.

Copyright © 2008 S. Karger AG, Basel

Introduction

The claudins are a family of highly related proteins that are important in tight junction formation and function. Claudin-1 was initially found in the liver of chickens [1]. Since that discovery 21 members of this family of proteins have been identified in humans [2]. *CLDN* gene expression is frequently altered in human cancers and claudin-1 expression has been found to be downregulated in various cancers [2]. Loss of function germline mutations in *CLDN1* were found in patients with neonatal ichthyosis-sclerosing cholangitis (NISCH) syndrome [3, 4].

Hepatitis C virus (HCV) is a blood borne infection with a worldwide distribution that varies markedly based on the risk of exposure to contaminated blood. It is estimated that about 130–170 million people have been infected with HCV. After initial infection with HCV, about 75–80% of people develop chronic infection, which is an

important cause of cirrhosis, end stage liver disease and hepatocellular carcinoma [5]. The age of HCV is not known. Determining when a virus originated is generally speculative and this is especially true for HCV, which was discovered only 20 years ago [6]. On the basis of the rate of sequence change HCV genotypes are estimated to have originated 500–2,000 years ago, however, this relatively short time period seems inconsistent with the widespread epidemiological distribution of HCV in human populations [7].

Recently it was shown that claudin-1, which is highly expressed in the liver [2], is required for HCV cell entry [8]. Residues within the first extracellular loop of claudin-1 are critical for viral entry and mutations in this region render cell-lines less permissive to HCV infection. Based on protein sequences, claudin-7 is the closest related member to claudin-1 [2]. It has been reported that claudin-7 may play a role in infection of CD4-negative cells by human immunodeficiency virus type 1 (HIV-1) [9].

Studies of *CCR5* $\Delta 32$, a loss of function mutation in the gene that encodes the primary HIV-1 co-receptor, have shed light on HIV-1 pathogenesis [10] and novel therapies that mimic this allele by blocking *CCR5* are effective against HIV-1 infection [11, 12]. Genetic variation in *CLDN1*, which encodes for a HCV viral receptor, could be associated with susceptibility, outcomes and treatment in HCV infection, but further studies are needed.

The requirement of claudin-1 for cell entry by HCV and the high worldwide prevalence of HCV infection raise the possibility that genetic variation in this gene could be under selective pressure of HCV or a related virus. The current tools of analyses of genetic variation in populations with different histories are well suited for determination of recent positive selection. With re-sequence data, it is also possible to look for indications of purifying selection in one or more populations.

We investigated common genetic variation in *CLDN1* in four self-described ethnically distinct U.S. populations to look for genetic variation and signatures of selection. Furthermore, we compared *CLDN1* with *CLDN7*, which is the most similar of the genes in the claudin family.

Patients and Methods

Population Samples

Re-sequencing analysis was performed on genomic DNA from the SNP500Cancer population, which consists of 102 unrelated individuals of self-described heritage: 24 of African/African American heritage (AA), 31 of Caucasian heritage (CA), 23 of Hispanic heritage (HI) and 24 of Pacific Rim heritage (PR) [13]

(<http://snp500cancer.nci.nih.gov>). With 204 chromosomes bi-directionally sequenced, the detection rate for SNPs with a minor allele frequency (MAF) >5% and >1% is 99% and 87%, respectively [14].

Detection of Variant Sites by Re-Sequencing of *CLDN1* and *CLDN7*

Primers were designed using the web-based version Primer3 software (<http://fokker.wi.mit.edu>) to cover the exons and evolutionary conserved regions within the gene. Sequence analysis was performed on amplified fragments generated by standard PCR techniques. 20 ng of genomic DNA was used per reaction and amplification conditions were as follows: initial 95°C for 10 min; 35 cycles at 94°C for 30 s, an assay specific optimized temperature between 54°C and 69°C for 45 s, 72°C for 45 s and a final extension at 72°C for 10 min. Bi-directional sequence analysis was performed using Dye Terminator method (ABI Perkin-Elmer, Foster City, Calif., USA) and was analyzed on ABI-Perkin Elmer platforms (model 3730) with Sequence Analysis 3.7 software and Seqscape software version 2.5 (Applied Biosystems).

The SNP frequencies were determined and haplotypes were estimated using PHASE version 2.1 [15]. DNAsp version 4.0 was used to estimate nucleotide diversity, as well as to identify synonymous and nonsynonymous substitutions [16]. Using this program we also calculated Tajima's D test statistic proposed by Tajima (1989), equation 38, for testing the hypothesis that all mutations are selectively neutral. The D test is based on the differences between the number of segregating sites and the average number of nucleotide differences. The confidence limits of D (two-tailed test) is obtained assuming that D follows the beta distribution (Tajima 1989, equation 47), i.e. the confidence limits given in table 2 of Tajima [17]. F_{st} was calculated as described by Hudson et al. [18]. Haploview version 3.31 (<http://www.broad.mit.edu/mpg/haploview/index.php>) was used to determine the degree of linkage disequilibrium (LD) and generate plots based on r^2 values [19]. Haplotype blocks were determined using the confidence intervals (CIs) set by Gabriel et al. [20] (upper CI = 0.98, lower CI = 0.7). HapMap Phase II data was used to look at LD in the regions flanking the genes, namely about 25 kb on each side [21]. The data were analyzed comparing the African, European and the combined Japanese and Chinese populations with each other. All SNPs and their frequencies in these populations can be found in Genewindow (<http://genewindow.nci.nih.gov>).

Claudin Homology and Phylogenetic Tree

Evolutionary conservation was estimated comparing 28 vertebrates (Human, Armadillo, Bushbaby, Cat, Chicken, Chimpanzee, Cow, Dog, Elephant, Frog, Fugu, Guinea pig, Hedgehog, Horse, Lizard, Medaka, Mouse, Opossum, Platypus, Rabbit, Rat, Rhesus, Shrew, Stickleback, Tenrec, Tetraodon, Tree shrew and Zebrafish) with the use of a phylogenetic hidden Markov model, phastCons [22] in the UCSC genome browser (<http://genome.ucsc.edu/cgi-bin/hgGateway>). This method calculates the posterior probability that each site was generated by the conserved state (Conservation score). Evolutionary conservation was also examined using the subset of 17 of these species that are placental mammals and, therefore, reflect more recent evolution. Evolutionary conserved regions were determined using a cut-off of 0.7 for the conserved areas.

Table 1. Nucleotide diversity and Tajima's D statistic for *CLDN1* and *CLDN7*

	Population				
	Total	AA	CA	HI	PR
A <i>CLDN1</i>					
Π	1.4×10^{-3}	1.5×10^{-3}	1.2×10^{-3}	1.4×10^{-3}	1.3×10^{-3}
Θ	1.7×10^{-3}	1.7×10^{-3}	1.1×10^{-3}	1.4×10^{-3}	1.2×10^{-3}
Tajima's D	-0.62 (NS)	-0.35 (NS)	0.37 (NS)	-0.07 (NS)	0.14 (NS)
B <i>CLDN7</i>					
Π	0.30×10^{-3}	0.18×10^{-3}	0.28×10^{-3}	0.30×10^{-3}	0.43×10^{-3}
Θ	0.43×10^{-3}	0.45×10^{-3}	0.32×10^{-3}	0.34×10^{-3}	0.34×10^{-3}
Tajima's D	-0.56 (NS)	-1.40 (NS)	-0.27 (NS)	-0.28 (NS)	0.56 (NS)

AA = African American; CA = Caucasian; HI = Hispanic; PR = Pacific Rim.

21 human claudin gene DNA sequences were downloaded from Ensemble release 45, June 2007 (<http://www.ensembl.org>). ClustalW software was used to produce a multiple sequence alignment and Jalview software was used to visualize the results. ClustalW phylogenetic calculations were based on the neighbor-joining method of Saitou and Nei [23].

Results

In the resequence analysis of *CLDN1*, which spans 16,723 bp on chromosome 3 (chr3: 191,506,187–191,522,909) and has 4 exons, 13,458 bp were analyzed: 12,602 bp (75%) of the genic region, plus 756 bp in the 5' untranslated region (UTR) and 100 bp in the 3' UTR. A total of 133 SNPs were identified, yielding an average of 1 SNP per 95 bp or 10.6 SNPs per kb. Of these SNPs, 57 had a MAF >5%. Six coding variants were observed of which two nonsynonymous SNPs, A124T and A135I, were each observed only once among the 102 individuals. In addition, there were 8 indels, 5 in intron 1 (MAFs, 1–11%), 1 in intron 3 (MAF, 1%) and 1 in a non-coding region of exon 4 (MAF, 1%). In the 5' UTR region an AC repeat length polymorphism was found 578 bp upstream from the start of exon 1.

The resequence analysis of *CLDN7* (chr17: 7,104,182–7,106,513), which has 60% sequence similarity to *CLDN1* and includes 2,331 bp, targeted 1,989 bp across 61% of the gene including the four exons, 468 bp of the 5' UTR and 103 bp of the 3' UTR. Only five SNPs were found in this region. The density of SNPs across *CLDN7* is 1 in 398 bp, which is substantially lower than that of *CLDN1*. One nonsynonymous SNP, A197V (MAF, 30% across all four populations) was observed. Additionally, one synonymous SNP and three SNPs outside the coding regions were detected. Two indels were found, one in the pro-

motor region of the gene (MAF, 10.3%) and one in the non-coding region of exon 1 (MAF, 40%).

Nucleotide Diversity and Genetic Heterozygosity

Table 1 shows the nucleotide diversity, which was calculated using the average number of nucleotide differences per site between two sequences (π) and the population mutation parameter (θ) for each gene within each subpopulation (African American, Caucasian, Hispanic and Pacific Rim), as well as overall ($n = 102$). For *CLDN1*, π was highest in African Americans (1.5×10^{-3}), but similar in the other subpopulations (1.2×10^{-3} to 1.4×10^{-3}). There was no difference in θ between the populations. Tajima's D statistic, which is based on the number of segregating sites, was calculated for *CLDN1* (table 1A). Tajima's D statistic was not significant for the four groups or overall.

The results for *CLDN7* are shown in table 1B. π was highest in the Pacific Rim and lowest in the African American group. θ did not show any difference between the populations and neither did Tajima's D.

We calculated the F_{st} statistic, a measure of population differentiation, to compare heterozygosity between the ethnic groups. For *CLDN1* the highest degree of heterozygosity ($F_{st} = 0.05$) was observed between Caucasian and African American populations and the lowest ($F_{st} = 0.005$) between Caucasians and Hispanics (table 2A). These data suggest no differentiation between the 4 populations. In comparison, the highest degree of heterozygosity for *CLDN7* was observed between African American and Pacific Rim populations ($F_{st} = 0.10$) and the lowest ($F_{st} = -0.02$) was between the Caucasian and Hispanic populations (table 2B).

Haplotypes and Linkage Disequilibrium

LD was determined by group and, when possible, haplotype blocks were estimated (fig. 1). Interestingly, neither gene displayed a significant degree of LD across the coding region of the gene. In fact, there is minimal LD across *CLDN1*, which has only one small block in the 3' end of the gene. Using the HapMap Phase II data we looked at LD across the wider gene region. An area of 50 kb with the gene in the middle was chosen (fig. 2). *CLDN1* lies in an area between two blocks of strong LD. Both the 5' UTR and the 3' UTR show blocks of LD in all three Hapmap populations (fig. 2A). *CLDN7* is found in a block of strong LD in the European population. LD disintegrates over a shorter distance in the two Asian populations combined, as well as in the African population (fig. 2B).

Conservation of Sequences between Species

The mean conservation score for *CLDN1* was 0.11 (SD: 0.28) for 28 vertebrate species and 0.13 (SD: 0.31) for 17 placental species. *CLDN7* had a mean conservation score of 0.31 (SD: 0.43) for the vertebrate species and 0.27 (SD: 0.42) for the placental species.

Apart from the coding regions, areas of *CLDN1* with conservation >0.7 were found in intron 1 and in the intronic areas around exons 3 and 4 for both the 28 vertebrate species (fig. 3A) and the 17 placental species (fig. 3B). This similarity suggests the preservation of sequences that could have functional activities. *CLDN7* shows similar conservation in intron 1, but not in the intronic areas around exons 3 and 4 (fig. 3C). As in *CLDN1*, most conservation in *CLDN7* seems to be old because there is little difference between all 28 vertebrate species and the subset of 17 placental species (fig. 3D).

ClustalW Comparison of DNA Sequence

Previously it was shown that the protein sequence of claudin-1 and claudin-7 cluster together within the protein family of claudins [2]. We compared the DNA sequences of the family of genes using ClustalW and found that *CLDN1* and *CLDN7* cluster together along with *CLDN19* (fig. 4).

Discussion

In a comparative re-sequence analysis of the two most closely similar genes in the claudin family, *CLDN1* and *CLDN7*, we observed no evidence for recent positive or negative selection. As expected, our re-sequence analysis

Table 2. F_{st} statistic for *CLDN1* and *CLDN7* in the four ethnic groups

A <i>CLDN1</i>		PR	HI	CA
	AA	0.03	0.01	0.05
	CA	0.04	0.005	
	HI	0.02		
B <i>CLDN7</i>		PR	HI	CA
	AA	0.10	0.08	0.08
	CA	0.003	-0.02	
	HI	-0.006		

AA = African American; CA = Caucasian; HI = Hispanic; PR = Pacific Rim.

revealed that the pattern of genetic diversity differs for these two genes, particularly in distinct populations, which we estimated by analyzing the four self-described populations of the SNP500Cancer set [13]. Moreover, neither gene shows any signatures of selection, and our sample size was sufficiently large to detect small signatures. Although several rare nonsynonymous variants are found in these genes, given the sample size, it is not possible to make a case for purifying selection. Like most genes studied to date, the analysis does not support a recent signature of positive selection and we conclude that *CLDN1*, which is a critical receptor for HCV, appears to be neutral with regard to selection.

Previously, loss of function mutations in *CLDN1* have been reported in children presenting with NISCH syndrome [3, 4]. In a Swiss girl with NISCH syndrome, a homozygous deletion in exon 2 caused a premature stopcodon and a truncated protein [3]. Four individuals with NISCH syndrome from 2 inbred Moroccan families shared a homozygous deletion in exon 1, which was absent in 52 unrelated Moroccan controls [4]. We found neither of these variants in our study population. When we examined the catalogue of genetic variants across *CLDN1*, we noted that the first extracellular loop of the protein, which appears to be crucial for HCV binding, did not reveal non-synonymous SNPs. Therefore, we observed no evidence for a common loss of function variant that could restrict disease, as has been established in HIV-1-infection for the *CCR5* $\Delta 32$ allele [10, 24, 25]. The population genetic parameters, estimated using our sequencing data for both *CLDN1* and *CLDN7*, appear to be consistent with being selectively neutral,

Fig. 1. Linkage disequilibrium plots showing r^2 for *CLDN1* of 4 ethnic subpopulations of the SNP500Cancer population. AA = African American, CA = Caucasian, HI = Hispanic, PR = Pacific Rim. Colors indicate the degree of LD. White: $r^2 = 0$, black: $r^2 = 1$, shades of grey: $0 < r^2 < 1$.

which argues against the selective pressure of HCV, a viral disease that has recently become common worldwide.

The nucleotide diversities for *CLDN1* (1.4×10^{-3}) and *CLDN7* (0.3×10^{-3}) fall well within the range found in other genes by direct sequencing. In a resequencing study of the T cell receptor, Mackelprang et al. found a

nucleotide diversity (π) of 12.7×10^{-4} [26]. In a study of 213 environmental response genes, π ranged from 0.72×10^{-4} to 19.3×10^{-4} [27] with the overall $\pi = 6.7 \times 10^{-4}$. And in a study using 180 SeattleSNPs genes comparing African American and European American populations, the former group was found to have a greater nucleotide diversity (9.01×10^{-4} and 6.96×10^{-4} ,

Color version available online

Color version available online

Fig. 2. LD plot of Phase II HapMap populations. Linkage Disequilibrium (LD) of *CLDN1* (**A**) and *CLDN7* (**B**) in three HapMap populations. The region comprises 50 kb of the area around the gene. Colors indicate the degree of LD. White: $r^2 = 0$, black: $r^2 = 1$, shades of grey: $0 < r^2 < 1$. The rectangle with the name of the gene above the diagrams indicates the size and position of the gene.

respectively) [28]. The difference in π between *CLDN1* and *CLDN7* might be explained by the larger non-coding region of *CLDN1*, or likely reflects the normal variance of genes, in other words, within the range for selectively neutral genes.

In a genome wide genotyping study of East Asian, African-American, and European-American individuals, the average F_{st} for autosomal SNPs was 0.107 in coding regions, 0.118 in intronic regions and 0.123 overall [29]. However, these estimates were based on selected genotyping of SNPs rather than resequencing as was done in our study. In our analysis of *CLDN1*, the highest F_{st} was 0.05 (for the comparison of African American and Caucasian populations). Because this is below what was found in genome wide analysis, it suggests no difference in selection between any of the populations studied. We found that *CLDN7* had a similar F_{st} with the highest value (0.1) for the comparison of the African American and Pacific Rim populations.

Various studies report on the estimates of Tajima's D. The mean value for Tajima's D in resequenced genes was estimated to be -0.54 for people of African descent and 0.26 for people of European descent [30]. This was substantially different from the 0.94 and 1.25 for the same groups as estimated from a dataset of dense SNP genotyping [30]. In 180 SeattleSNPs genes the average Tajima's D in African-American was found to be -0.51 and 0.18 in European Americans [28]. In our study, Tajima's D for the 4 populations combined was -0.6 for both *CLDN1* and *CLDN7*. As was shown in a resequencing study of 132 genes [31], this does not represent evidence for recent selective pressure.

The DNA sequences of *CLDN1* and *CLDN7* cluster together with *CLDN19*. These 3 genes also cluster together based on protein sequences, as reported in the past [32]. *CLDN1* is about 7.3 times larger than *CLDN7*, although both genes encode a protein of 211 amino acids. Despite these differences in intronic region these genes cluster together in the ClustalW analyses of both protein and DNA sequences within the claudin family of genes, suggesting a common ancestor gene.

For both *CLDN1* and *CLDN7*, a high degree of conserved sequence in intron 1 was observed among 28 different vertebrate species, which is expected because of the putative regulatory sequences in the first intron. The genes also show a high degree of similarity in protein and DNA sequence, as expected in a family of genes with overlapping functional attributes in tight junction formation.

Fig. 4. Clustal W analyses of claudin-gene family members. **A** Phylogram of ClustalW comparison of DNA sequence of the members of the *CLDN* family. **B** Cladogram of ClustalW comparison of DNA sequence of the members of the *CLDN* family.

Recently it was shown that alignment analysis has uncertainties and that different programs can lead to different conclusions [33]. Interpretation of our data should be done with this caution in mind. The same is true for summary statistics such as Tajima's D [31], as population demography and the recombination rate can influence the outcome of this test. In this study we resequenced the 2 genes in 102 individuals. With this approach, we estimate a 99% probability to find a segregating site that in a larger data set would be estimated to have a MAF over 5%. However, less common SNPs have a lower change to be

found, which might influence our results. Also the use of HapMap data has its limitations especially the use of haplotype estimates [34].

In conclusion, we conducted a resequence analysis of claudin genes implicated in HCV entry and found no evidence of positive selection, thus suggesting that the genes are most likely selectively neutral. In this regard, functional mutations were not found in otherwise healthy individuals, suggesting a degree of constraint on these two genes, of which one is a late stage receptor for HCV infection.

Acknowledgments

We would like to thank Renee Chen, Nicholas Orr, Hye Kim and Jun Fang for their technical assistance, as well as Charles Rice and Thomas von Hahn for helpful conversations about claudin-1.

Fig. 3. Interspecies conservation assessment by Phastcons algorithm. **A** *CLDN1* compared between 28 vertebrate species. **B** *CLDN1* compared between a subset of 17 placental species. **C** *CLDN7* compared between 28 vertebrate species. **D** *CLDN7* compared between a subset of 17 placental species. The horizontal line indicates a conservation score of 0.7.

References

- Furuse M, Fujita K, Hiiragi T, Fujimoto K, and Tsukita S: Claudin-1 and -2: novel integral membrane proteins localizing at tight junctions with no sequence similarity to occludin. *J Cell Biol* 1998;141:1539–1550.
- Hewitt KJ, Agarwal R, Morin PJ: The claudin gene family: expression in normal and neoplastic tissues. *BMC Cancer* 2006;6:186.
- Feldmeyer L, Huber M, Fellmann F, Beckmann JS, Frenk E, Hohl D: Confirmation of the origin of NISCH syndrome. *Hum Mutat* 2006;27:408–410.
- Hadj-Rabia S, Baala L, Vabres P, et al: Claudin-1 gene mutations in neonatal sclerosing cholangitis associated with ichthyosis: a tight junction disease. *Gastroenterology* 2004;127:1386–1390.
- Lauer GM, Walker BD: Hepatitis C virus infection. *N Engl J Med* 2001;345:41–52.
- Choo QL, Kuo G, Weiner AJ, Overby LR, Bradley DW, Houghton M: Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science* 1989;244:359–362.
- Simmonds P: Reconstructing the origins of human hepatitis viruses. *Phil Trans R Soc Lond B Biol Sci* 2001;356:1013–1026.
- Evans MJ, von Hahn T, Tscherne DM, et al: Claudin-1 is a hepatitis C virus co-receptor required for a late step in entry. *Nature* 2007;446:801–805.
- Zheng J, Xie Y, Campbell R, et al: Involvement of claudin-7 in HIV infection of CD4(–) cells. *Retrovirology* 2005;2:79.
- Dean M, Carrington M, Winkler C, et al: Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CCR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. *Science* 1996;273:1856–1862.
- Lalezari J, DeJesus E, Lampiris H, et al: Efficacy and safety of maraviroc plus optimized background therapy in viremic ART-experienced patients infected with CCR5-tropic HIV-1: 24 week results from a phase 2b/3 study in the US and Canada. Fourteenth Conference on Retroviruses and Opportunistic Infections, Los Angeles, abstract 104bLB, 27th February 2007.
- Nelson M, Fätkenheuer G, Konourina I, et al: Efficacy and safety of maraviroc plus optimized background therapy in viremic ART-experienced patients infected with CCR5-tropic HIV-1 in Europe, Australia and North America: 24 week results. Fourteenth Conference on Retroviruses and Opportunistic Infections, Los Angeles, abstract 104aLB, 2007.
- Packer BR, Yeager M, Burdett L, et al: SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. *Nucleic Acids Res* 2006;34: D617–D621.
- Kruglyak L, Nickerson DA: Variation is the spice of life. *Nat Genet* 2001;27:234–236.
- Stephens M, Donnelly P: A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 2003;73:1162–1169.
- Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R: DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 2003;19:2496–2497.
- Tajima F: Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989;123:585–595.
- Hudson RR, Slatkin M, Maddison WP: Estimation of levels of gene flow from DNA sequence data. *Genetics* 1992;132:583–589.
- Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005;21: 263–265.
- Gabriel SB, Schaffner SF, Nguyen H, et al: The structure of haplotype blocks in the human genome. *Science* 2002;296:2225–2229.
- Frazer KA, Ballinger DG, Cox DR, et al: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449:851–861.
- Siepel A, Bejerano G, Pedersen JS, et al: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005;15:1034–1050.
- Saitou N, Nei M: The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;4: 406–425.
- Ioannidis JP, Rosenberg PS, Goedert JJ, et al: Effects of CCR5-Delta32, CCR2–64I, and SDF-1 3'A alleles on HIV-1 disease progression: An international meta-analysis of individual-patient data. *Ann Intern Med* 2001; 135:782–795.
- Ioannidis JP, Contopoulos-Ioannidis DG, Rosenberg PS, et al: Effects of CCR5-delta32 and CCR2-64I alleles on disease progression of perinatally HIV-1-infected children: an international meta-analysis. *AIDS* 2003;17: 1631–1638.
- Mackelprang R, Livingston RJ, Eberle MA, et al: Sequence diversity, natural selection and linkage disequilibrium in the human T cell receptor alpha/delta locus. *Hum Genet* 2006; 119:255–266.
- Livingston RJ, von Niederhausern A, Jegga AG, et al: Pattern of sequence variation across 213 environmental response genes. *Genome Res* 2004;14:1821–1831.
- Crawford DC, Akey DT, Nickerson DA: The patterns of natural variation in human genes. *Annu Rev Genomics Hum Genet* 2005;6: 287–312.
- Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 2002;12:1805–1814.
- Carlson CS, Thomas DJ, Eberle MA, et al: Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res* 2005;15:1553–1565.
- Akey JM, Eberle MA, Rieder MJ, et al: Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol* 2004;2:e286.
- Kramer F, White K, Kubbies M, Swisshelm K, Weber BH: Genomic organization of claudin-1 and its assessment in hereditary and sporadic breast cancer. *Hum Genet* 2000; 107:249–256.
- Wong KM, Suchard MA, Huelsenbeck JP: Alignment uncertainty and genomic analysis. *Science* 2008;319:473–476.
- Zhao H, Pfeiffer R, Gail MH: Haplotype analysis in population genetics and association studies. *Pharmacogenomics* 2003;4: 171–178.