Mammalian Genome 13, 554–557 (2002).
DOI: 10.1007/s00335-002-2169-4

Mammalian Genome
Incorporating Mouse Genome

SNPs in putative regulatory regions identified by human mouse comparative sequencing and transcription factor binding site data

Poulabi Banerjee,[1] Melanie Bahlo,[2] Jody R. Schwartz,[1] Gabriela G. Loots,[1] Kathryn A. Houston,[1] Inna Dubchak,[1,3] Terence P. Speed,[2] Edward M. Rubin[1]

[1]Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, California, 94720, USA
[2]Division of Genetics and Bioinformatics, The Walter & Eliza Hall Institute of Medical Research, Parkville, VIC 3050, Australia
[3]National Energy Research Supercomputing Center, Lawrence Berkeley National Laboratory, Berkeley, California, 94720, USA

Received: 29 January 2002 / Accepted: 22 May 2002

Genome-wide disease association analysis by using SNPs is being explored as a method for dissecting complex genetic traits (Risch and Merikangas 1996), and a vast number of SNPs have been generated for this purpose. As there are cost and throughput limitations of genotyping large numbers of SNPs and statistical issues regarding the large number of dependent tests on the same data set, to make association analysis practical it has been proposed that SNPs should be prioritized based on likely functional importance (Risch 2000). The most easily identifiable functional SNPs are coding SNPs (cSNPs), and accordingly cSNPs have been screened in a number of studies (Cargill et al. 1999; Halushka et al. 1999). SNPs in gene regulatory sequences embedded in noncoding DNA are another class of SNPs suggested for prioritization owing to their predicted quantitative impact on gene expression. The main challenge in evaluating these SNPs, in contrast to cSNPs, is a lack of robust algorithms and databases for recognizing regulatory sequences in noncoding DNA. Approaches that have been previously used to delineate noncoding sequences with gene regulatory activity include cross-species sequence comparisons and the search for sequences recognized by transcription factors (Gottgens et al. 2000; Loots et al. 2000). We combined these two methods to sift through mouse human genomic sequences to identify putative gene regulatory elements and subsequently localized SNPs within these sequences in a 1-Megabase (Mb) region of human Chromosome (Chr) 5q31, orthologous to mouse Chr 11 containing the Interleukin cluster (Fig. 1a).

To survey SNPs in conserved noncoding sequences (CNS), we screened 40 individuals of European ($n = 23$) and African ($n = 17$) origin. A CNS was selected by cross-species sequence alignments with an inclusion criterion of ≥80% identity over ≥120bp by intersection/union analyses of a 200-kb region of this interval (Dubchak et al. 2000). CNS ($n = 52$) meeting these criteria (Fig. 1a) were screened, and of these CNS1 (Loots et al. 2000), CNS 7 (Cockerill et al. 1999), CNS23 (Cockerill et al. 1999), CNS2 and CNS17 (G.G. Loots personal communication) have been experimentally verified as gene regulatory elements. In addition to the CNS, two genes in the interval, *IL3* and *CSF2,* were also sequenced, and we obtained a total of 85 SNPs, of which 28 were located in the CNS. Characteristics of the SNPs, categorized by location in coding, non-coding, CNS or intergenic sequences are provided in Table 1a.

For a comparative study of sequence diversity, in this 1-Mb region, we estimated the average heterozygosity (H) and the segregating sites parameter (θ_S). Under the assumptions of neutrality, Hardy-Weinberg equilibrium (HWE) and a mutation model that assumes only one mutation at each bp, (known as the infinite sites model) (Li 1997), θ_S is also an estimator of the mutation rate. The total levels of nucleotide diversity (Table 1a) in this 1-Mb region are consistent with those observed in previous studies (Halushka et al. 1999; Sunyaev et al. 2000). SNPs were found at a similar frequency of 1 per 494 bp in coding sequences with $\theta_S = 4.09 \times 10^{-4}$, and 1 per 403 bp in the CNS with $\theta_S = 5.01 \times 10^{-4}$. The similar values of average heterozygosity (H) and the theta estimator derived from the segregating sites for the various classes of SNPs (Table 1a) indicate that the standard neutral model is applicable over this region as a whole. However, this does not preclude selection acting on parts of this genomic region, such as the exons of *CSF2*. Since in this sample there are few SNPs within these genes, there is a lack of power to detect departures from neutrality by statistical tests based on summary measures such as the average heterozygosity and θ_S based merely on sequence data, i.e., non-phased data.

Allele frequencies at SNPs have implications for the sample size required for disease association studies and are important in defining population sub-structures as well. The frequency spectra of the minor allele for the SNPs are shown in Fig. 1b. The frequency distribution of SNPs specific to individuals of African (41) and Northern European (14) descent, as well as the ones shared (30), is also depicted in Fig. 1b. The African samples produced more SNPs than the Northern European samples, and this is also reflected in the estimates of nucleotide diversity (Table 1b), across all categories. Our results agree with previous studies and demonstrate greater diversity and more unique SNPs in the African population.

The two samples were tested for differences between the two populations by using the two summary measures, H and θ_S, by using a permutation test. Accurate empirical distributions using two sample *t*-test-like statistics for both estimators were generated with 5000 random permutations of the data of a possible 1.52×10^{11} permutations. The *t*-statistic based on θ_S was not a useful discriminant between the two populations, resulting in a spiky, multimodal empirical distribution for the statistic. Table 1a contains the *p*-values generated for the statistics based on the average heterozygosity. After adjustment for multiple testing (Bonferroni correction, 10 tests) the threshold for significance for a single test becomes $p = 0.005$. As expected, exons showed no discernible differences between the two samples; however, CNS regions showed a significant difference.

Linkage Disequilibrium (LD) was examined for the region by taking advantage of the Expectation-Maximization (EM)

Current address of P. Banerjee: Genomic and Proteomic Sciences, Pfizer Global Research and Development, Groton, CT 06340, USA
Correspondence to: E.M. Rubin; E-mail: EMRubin@lbl.gov

algorithm implemented as a permutation test in the software ARLEQUIN (Schneider et al. 2000). LD can be estimated only from haplotype data. The EM algorithm (Slatkin and Excoffier, 1996) allows the estimation of haplotypes and their frequencies, thus providing an opportunity to examine LD with sequence data. Details of the procedure can be found in the ARLEQUIN manual. The LD was examined for the combined data from both populations by using a sliding window approach within pairs of SNPs so that each SNP overlapped with another. Thus, each test window was moved along by one SNP, and the next pair of SNPs was tested for LD; e.g., the first data point on Fig. 1c is the LD between the first SNP pair (SNP pair 1 and 2), whereas the second data point is the LD between the second SNP pair (SNP pair 2 and 3), and this sequence was employed for all SNPs to determine LD across the region. In this sample of Northern Europeans and Africans, high LD was observed around the coding regions, *AF5Q31*, *Cyclin1-homolog*, *IL4*, *and IL3*, with the highest around the *SEPT8* gene (Fig. 1c).

Population-specific SNPs in general have an inverse-J shaped distribution with the majority of sites (96%) having a minor allele frequency of less than 10%. Under the assumption that these SNPs have arisen only once and that they are neutral, these low frequencies would indicate the relative youth of these polymorphisms. The average allele frequency (minor allele) for population-specific SNPs is 3.3%, while that for shared SNPs is 30.4%. SNPs in CNS are uncommon, with 54% having a minor allele frequency less than 10%.

To further prioritize SNPs present in CNS, we ascertained sites that fall in predicted transcription factor binding sites (TFBS) conserved between mice and humans. Our method employed a computational tool, *rVISTA*, for high-throughput discovery of *cis*-regulatory elements that combines clustering of predicted TFBS and the analysis of inter-species sequence conservation to maximize the identification of functional sites (Loots et al. 2002). The potential of *rVISTA* to identify true positive TFBS while reducing the prediction of false positives was analyzed across this 1-Mb sequence by using a number of AP-1, NFAT, and GATA-3 sites that have been experimentally verified. By employing the orthologous human–mouse dataset, >95% of the ~58,000 TFBS predicted on the human sequence were eliminated, and *rVISTA* identified 88% of the experimentally verified AP-1, NFAT, and GATA-3 sites (Loots et al. 2002). We employed *rVISTA* to identify putative TFBS across this region and localized 25 of the SNPs generated in CNS within these sites. Of these 25 SNPs, 19 were located in the four base pair core of the putative TFBS, indicating that these SNPs might be functionally important (Table 1c).

In this 1-Mb region, five (10%) of the CNS identified have been experimentally shown to impact on gene expression, whereas only one (2%) known silencer found in the 3′UTR of the *IL4* gene (Kubo et al. 1997) was not identified by human/mouse sequence conservation. The observation that approximately 10% of the CNS have been experimentally verified supports our sequence-based strategy for defining potential gene regulatory elements. Of the 28 SNPs discovered in the CNS, 25 can be further prioritized as gene regulatory SNPs based on their location in evolutionarily conserved TFBS. This study, using currently available (human and mouse genomic sequence) databases, identified 25 selectively neutral, putative regulatory SNPs. These SNPs, localized to a 1-Mb region of the genome to which several common quantitative

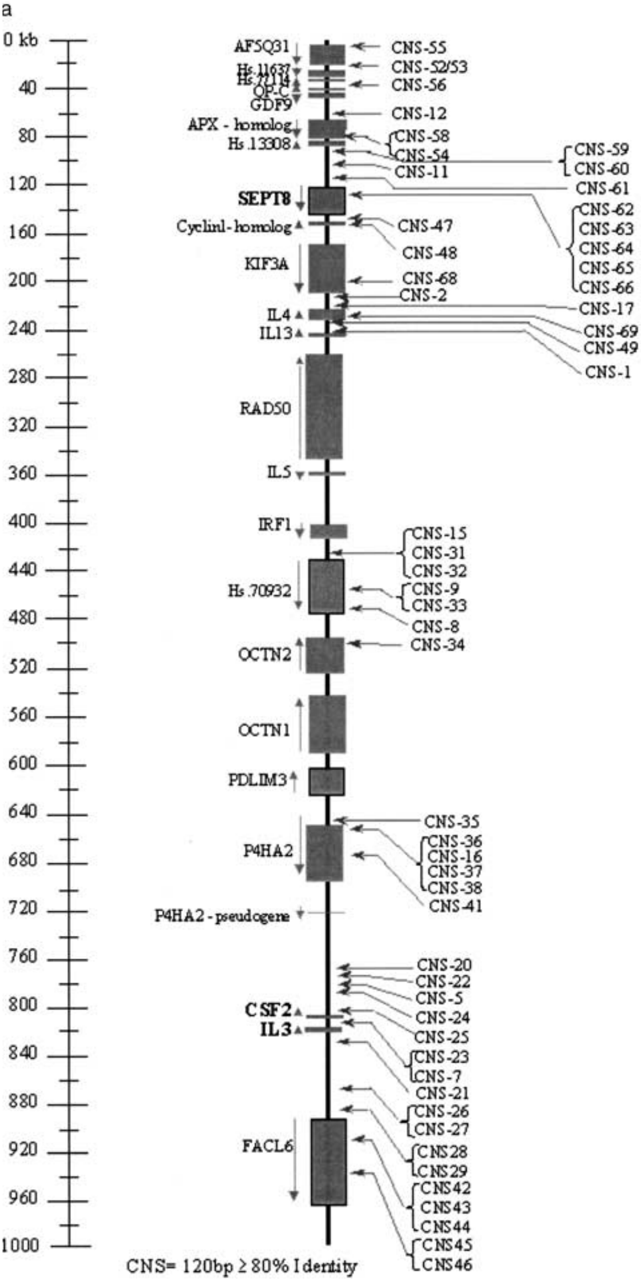

Fig. 1. Putative gene regulatory SNPs of 1Mb of human 5q31. **(a)** Map of the 1-Mb region depicting the genes in the region and the 52 highly conserved non-coding sequences (labeled arrows) and the genes (in bold) *CSF2*, *IL3*, and the *SEPT8* used to survey for SNPs. **(b)** Distribution of minor allele frequency of SNPs categorized by the presence in individuals of African or Northern European descent (population-specific) or their presence in both (shared). **(c)** Plot of the S statistic $\{S = -2 \log\ (L_{H*}/L_H)\}$ versus pairwise SNP combinations moving along the entire length of the 1 Mb of human 5q31. The SNP data were partitioned into fragments containing three and nine SNPs, and for each fragment, pairwise comparisons of adjacent SNPs were calcluated by using ARLEQUIN. (All pairwise comparisons were not calculated owing to computational constraints.) The EM algorithm assumes Hardy-Weinberg equilibrium for the SNP allele frequencies, and the rejection of the test could also be due to departure from Hardy-Weinberg equilibrium. (*Continued on next page.*)

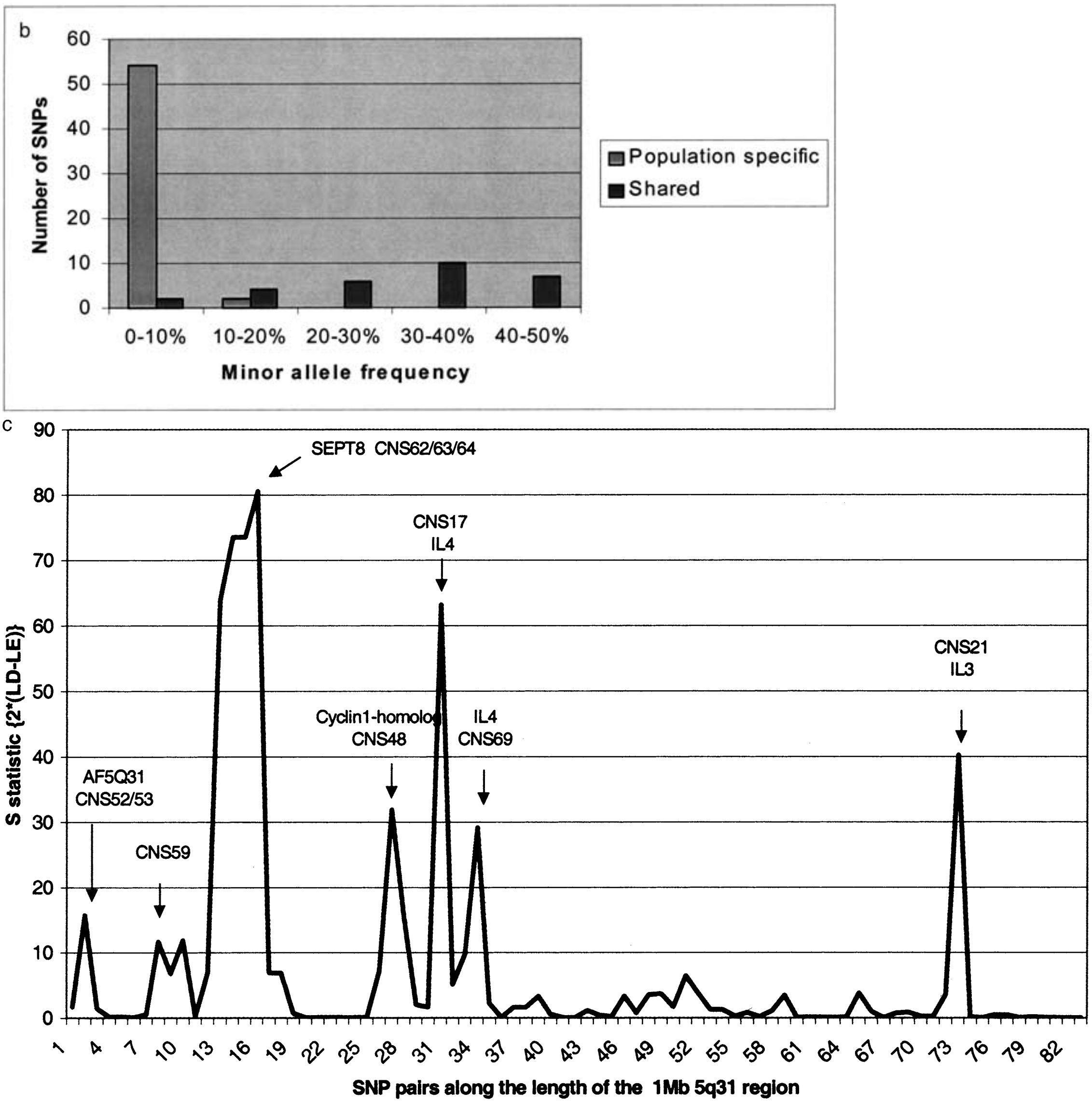

Fig. 1. *Continued.*

Table 1. Characteristics of the SNPs surveyed. (a) Amount of sequence screened and the number of SNPs generated for Chr 5q31. The nucleotide diversity ($\theta_S \times 10^{-4}$) and the permutation test results are also given for the various sequence categories. The permutation test was carried out within the various categories of SNPs and between the two populations of Africans and Northern Europeans.

Categories	Sequence screened (bp)	Number of SNPs	Frequency SNP/bp)	θ_S($\times$ 10^{-4})	Permutation Test	
					t-statistic	*p*-value
Coding	1.977	4	1/494	4.1 ± 2.2	−1.0	ns[a]
Intron	14.700	40	1/367	5.6 ± 1.6	−13.5	0.007
UTR	1.105	4	1/276	7.3 ± 4.0	−1.2	ns
CNS	11.295	28	1/403	5.0 ± 1.5	−9.9	0.002
Intergenic	13.996	36	1/389	5.2 ± 1.5	−9.9	0.038
Total	31.778	85	1/374	5.4 ± 1.5	−16.1	0.011

[a] ns, not siginificant

disorders have been mapped [e.g., asthma (Marsh et al. 1994), inflammatory bowel disease (Rioux et al. 2000)], are a useful resource for direct-association or linkage-disequilibrium analyses.

Acknowledgments. This work was supported by the SNP component (401HL66691) of the Program for Genomic Applications (PGA) funded by the National Heart, Lung, and Blood Institute (NHLBI/ NIH). Research was conducted at the Lawrence Berkeley National

Table 1b. Nucleotide diversity ($\theta_S \times 10^{-4}$) comparisons between African and Northern European samples.

	Populations			
	Africans/African–Americans		Northern Europeans	
Categories	Number of SNPs	$\theta_S(\times 10^{-4})$	Number of SNPs	$\theta_S(\times 10^{-4})$
Coding	2	2.5 ± 1.8	3	3.5 ± 2.1
Intron	34	5.7 ± 1.9	22	3.4 ± 1.2
Intergenic	31	5.4 ± 1.9	17	2.8 ± 1.0
UTR	3	6.6 ± 4.2	2	4.1 ± 3.0
CNS	24	5.2 ± 1.8	18	3.6 ± 1.3

Table 1c. SNPs located within the four base pair core of predicted evolutionarily conserved transcription factor binding sites. The SNP number refers to the position on the 1Mb sequence and the SNP is indicated in bold in the consensus sequence.

SNP Location	SNP Alleles (Major/Minor)	Transcription Factor	Consensus Sequence
CNS56	T/A	OCT1	aaagCAA**T**tacac[1,2]
CNS56	T/A	CDXA	a**T**TTGct[3]
CNS62	C/T	CAP	aC**A**GTatti[1,2]
CNS64	A/G	CAP	gCAC**A**cag
CNS65	C/T	AML1	cgC**G**GT[1, 3]
CNS65	T/C	TH1E47	cctctgctC**T**GGctat[3]
CNS65	T/C	CEBP	gcTC**T**Ggctattt[3]
CNS65	T/C	NF1	ctc**T**GGCtatttaaggcc[3]
CNS48	C/G	RORA1	ccttgtg**G**GTCac[2]
CNS68	G/A	CP2	gcaagatCAA**G**[3]
CNS2	C/G	GR	agacaatctgaT**G**TTctga[1,2]
CNS2	C/G	MYB	cagAA**C**Atca[2]
CNS2	C/G	CAP	a**C**ATCaga[2]
CNS17	T/C	OCT1	gacaG**A**ATggtaa[1,3]
CNS17	T/C	CDXA	cAT**T**Ctg[3]
CNS15	A/G	CAP	aC**A**TCtca[3]
CNS9	C/T	CEBPB	ggt**T**TGCagcattg[1,2]
CNS8	T/C	IK2	tgca**T**GGAttcc
CNS24	C/T	CETS1P54	ctgca**G**GAGggag[1,3]
CNS24	G/A	CEBP	gatT**T**GGgaaccat[1,2]
CNS24	G/A	CDXA	aTT**T**Ggg[1,2]
CNS23	T/A	DELTAEF1	aactACC**T**gaa[3]
CNS7	A/C	CETS1P54	agacaGG**A**Gaaaa[3]
CNS7	A/C	HSF1	**A**GAAaaccct
CNS21	C/T	ELK1	accagc**G**GAAatacaa[1,3]
CNS21	C/T	CETS1P54	ccagc**G**GAAatac[1,3]
CNS27	A/G	CAP	c**C**AACtta[1,2]
CNS29	A/G	HNF3B	atc**T**CTTtgcaa[1,3]

[1] The predicted transcription factor binding site is on the opposite DNA strand. The transcription factor binding site consensus sequence with the [2] major allele and [3] minor allele falls below the cutoff core similarity value of 0.75.

Laboratory (Dept. of Energy Contract DE-AC0376SF00098), University of California, Berkeley.

References

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K et al. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22, 231–238

Cockerill PN, Bert AG, Roberts D, Vadas MA (1999) The human granulocyte-macrophage colony-stimulating factor gene is autonomously regulated in vivo by an inducible tissue-specific enhancer. Proc Natl Acad Sci USA 96, 15097–15102

Dubchak I, Brudno M, Loots GG, Pachter L, Mayor C et al. (2000) Active conservation of noncoding sequences revealed by three-way species comparisons. Genome Res 10, 1304–1306

Gottgens B, Barton LM, Gilbert JG, Bench AJ, Sanchez MJ et al. (2000) Analysis of vertebrate SCL loci identifies conserved enhancers. Nat Biotechnol 18, 181–186

Halushka MK, Fan JB, Bentley K, Hsie L, Shen N et al. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 22, 239–247

Kubo M, Ransom J, Webb D, Hashimoto Y, Tada T et al. (1997) T-cell subset-specific expression of the IL-4 gene is regulated by a silencer element and STAT6. EMBOJ 16, 4007–4020

Li W-H (1997) *Molecular Evolution*. (Sunderland, Mass.: Sinauer Associates, Inc.)

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W et al. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288, 136–140

Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM (2002) rVista for comparative sequence based discovery of functional transcription Factor binding sites. Genome Res 12, 832–839

Marsh DG, Neely JD, Breazeale DR, Ghosh B, Freidhoff LR et al. (1994) Linkage analysis of IL4 and other chromosome 5q31.1 markers and total serum immunoglobulin E concentrations. Science 264, 1152–1156

Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, McLeod RS et al. (2000) Genomewide search in Canadian families with inflammatory bowel disease reveals two novel susceptibility loci. Am J Hum Genet 66, 1863–1870

Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405, 847–856

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273, 1516–1517

Schneider S, Roessli D, Excoffier L (2000) Arlequin: a software for population genetics data analysis. Ver 2.000., 2.0 edn. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva

Sunyaev SR, Lathe WC, Ramensky VE, Bork I (2000) SNP frequencies in human genes an excess of rare alleles and differing modes of selection. Trends Genet 16, 335–337