Annu. Rev. Genomics Hum. Genet. 2005. 6:287–312
doi: 10.1146/annurev.genom.6.080604.162309

First published online as a Review in Advance on May 19, 2005

THE PATTERNS OF NATURAL VARIATION IN HUMAN GENES

Dana C. Crawford, Dayna T. Akey, and Deborah A. Nickerson

Department of Genome Sciences, University of Washington, Seattle, Washington 98195; email: dcrawfo@gs.washington.edu, dakey@u.washington.edu, debnick@u.washington.edu

Key Words SNPs, haplotypes, diversity, association, linkage disequilibrium, Environmental Genome Project, SeattleSNPs Program for Genomic Applications

■ **Abstract** Currently, more than 10 million DNA sequence variations have been uncovered in the human genome. The most detailed variation discovery efforts have focused on candidate genes involved in cardiovascular disease or in susceptibilities associated with exposure to environmental agents. Here we provide an overview of natural genetic variation from the literature and in 510 human candidate genes resequenced for variation discovery. The average human gene contains 126 biallelic polymorphisms, 46 of which are common (≥5% minor allele frequency) and 5 of which are found in coding regions. Using this complete picture of genetic diversity, we explore conservation, signatures of selection, and historical recombination to mine information useful for candidate gene association studies. In general, we find that the patterns of human gene variation suggest that no one approach will be appropriate for genetic association studies across all genes. Therefore, many different approaches may be required to identify the elusive genotypes associated with common human phenotypes.

INTRODUCTION

Since the completion of the sequencing of the human genome (46, 108), there have been great strides in cataloguing and describing human gene variation (e.g., 83). The most common genetic variation in the human genome is the single nucleotide polymorphism (SNP). Currently, there are more than 10 million SNPs recorded in dbSNP (build 123), the public repository for DNA variations (see Electronic Databases). With a complete catalogue of common SNPs nearly in sight, investigators are now developing methods to apply SNPs to genetic association studies with the hopes of identifying DNA variants that contribute to an increased susceptibility to human diseases. With this in mind, we use this review to not only describe the patterns of human gene variation, but also to practically demonstrate how this variation may be applied to the next generation of well-designed genetic association studies.

SINGLE NUCLEOTIDE POLYMORPHISMS IN HUMAN GENES

Resequencing genes or genomic regions in several population samples is considered the gold standard for cataloguing genetic variation for any species. Currently, more than 11.5 megabases of reference sequence have been scanned for DNA variation discovery by two of the largest targeted resequencing efforts undertaken: the National Heart, Lung, and Blood Institute's Program for Genomic Applications and Pharmacogenetics and Risk of Cardiovascular Disease (SeattleSNPs PGA and PARC, collectively referred to here as SeattleSNPs PGA) and the National Institute for Environmental Health Sciences' Environmental Genome Project (EGP). If the number of individuals resequenced by these two projects is taken into account, close to 1 gigabase has been resequenced, which is equivalent to one third of the human genome and underscores the dramatic leaps in DNA sequencing technology. The SeattleSNPs PGA project resequences genes involved in inflammation, lipid metabolism, and blood pressure regulation (10, 26), whereas the EGP project resequences genes involved in DNA repair, cell cycle regulation, drug metabolism, and apoptosis (58). For most genes less than 35 kb in length, the entire genomic transcript, including all exons and introns, as well as ~2 kb upstream of the gene and ~1.5 kb downstream of the gene, is targeted for resequencing. For larger genes (more than 35 kb in length), introns are not typically completely sequenced, but are "sampled." These two projects differ from other large resequencing efforts (7, 84, 96) in that both coding and noncoding regions (including introns) of the gene are targeted in resequencing for variation discovery. A detailed description of the laboratory methods for these projects is available on the SeattleSNPs PGA and EGP Web sites (see Electronic Databases).

To date, the SeattleSNPs PGA has resequenced approximately 3.8 megabases across 180 genes in 47 individuals, comprising 23 Americans of European descent and 24 Americans of African descent. The EGP has resequenced 7.7 megabases across 330 genes in a subset of 90 individuals from the Polymorphism Discovery Resource 450 Panel (20), a panel designed to represent the American population with 24 Americans of European descent, 24 Americans of African descent, 24 Asians, 12 Hispanics, and 6 Native Americans. More than 75% of the targeted transcript was resequenced for approximately 70% of the genes presented here. The average candidate gene size is 20.9 kb in the SeattleSNPs PGA data set and 44.6 kb in the EGP data set. Together, these two projects have annotated 64,451 DNA sequence variations in 510 candidate genes spanning all 23 human chromosomes. The 510-gene data set represents approximately 2% of the predicted protein coding genes in the human genome (47). As expected, for the SeattleSNPs PGA data set, the African-American population has a greater overall number of variants (17, 701) compared with the European-American sample (11,009). The average nucleotide diversity (π) per gene is 9.01×10^{-4} and 6.97×10^{-4} for the African- and European-American samples, respectively. This measure of nucleotide diversity predicts one SNP every 1110 basepairs (bp) (African-Americans) and 1435 bp (European Americans) when any two chromosomes are compared. All DNA variations identified by the SeattleSNPs PGA and EGP, as well as allele frequencies

and individual genotypes, are routinely deposited in GenBank and dbSNP, the public repository for DNA variations (91, 92, 113). All DNA variations and corresponding data are also available for viewing and/or download on our Web sites (see Electronic Databases), and the list of genes analyzed here can be requested from the authors.

DNA sequence variations annotated by these two projects consist of both SNPs and insertion/deletion polymorphisms (3), collectively referred to as SNPs here. The occurrence of SNPs across the human genome is common, with one approximately every 180 bp, based on resequencing 137 individuals. The estimate from these data agrees with estimates derived from population genetics theory (55) and from empirical estimates based on other resequencing surveys (43, 96). Although the occurrence of SNPs across the genome is frequent, we and other investigators find that most SNPs are rare (64% of all SNPs), with a minor allele frequency (MAF) of <5%. Also, fewer SNPs are within coding regions of the gene [coding SNPs (cSNPs) 4%] compared with noncoding regions of the gene (96%), as previously observed (42, 58). The average gene in the combined SeattleSNPs PGA and EGP data sets (such as *VCAM1*) (Figure 1) has 126 SNPs, 46 of which are common (MAF $\geq$ 5%) and 5 of which are cSNPs. These observations have major implications for performing candidate gene association studies using either direct or indirect approaches, both of which we discuss further below.

ASSOCIATION STUDIES AND SINGLE NUCLEOTIDE POLYMORPHISMS

Association studies have become the focus of most study designs for identifying loci that are involved in complex, common human disease (e.g., heart disease, stroke, diabetes, and cancer). Prior to the completion of the Human Genome Project and the emergence of dense genetic maps, investigators used linkage studies and positional cloning to identify DNA mutations that caused rare disorders such as cystic fibrosis (52, 81) and Huntington's disease (40, 104). Despite the success of identifying genes segregating in classic Mendelian fashion as recessive or dominant disorders (37), investigators have been less successful in identifying loci that contribute to complex, common diseases. A landmark paper in 1996 by Risch & Merikangas (82) suggested that association study designs could be more powerful compared with linkage study designs in identifying the elusive susceptibility loci that geneticists seek. Furthermore, it was suggested that common DNA variation, as opposed to rare mutations, could be responsible for a proportion of common human diseases [i.e., the common variant/common disease (CV/CD) hypothesis] (14, 22, 56). Even though these suggestions and their potential for success in identifying genetic loci involved in common human diseases remain controversial (112), resources for association studies, such as dense genetic maps of SNPs across the human genome, are enabling investigators to more rapidly identify disease-causing loci that could potentially have a major impact on public health (21).

THE "DIRECT" APPROACH

As investigators begin to mine the resources available for association designs, two approaches have emerged for studying candidate genes: direct and indirect (22). In a direct candidate gene association study, the putative causative SNP is genotyped directly. The challenge of this approach is predicting or determining a priori which SNPs are likely to be causative or predict the phenotype of interest. Oftentimes, a SNP is suspected to be causative if it is a nonsynonymous SNP (nsSNP); that is, if the cSNP changes the amino acid in the protein of the gene of interest (reviewed in 5). For example, in a simple yet elegant experiment, Cohen et al. (19) hypothesized that genetic variants in three candidate genes are responsible for very low levels of high-density lipoprotein cholesterol (HDL-C), and that these variants would be found more often in individuals with low HDL-C compared with individuals with high HDL-C. Cohen et al. (19) resequenced the coding regions and splice sites of three candidate genes within individuals with extremely low and high levels of HDL-C and identified several nsSNPs over-represented among individuals with low HDL-C compared with high HDL-C. In contrast to the positive finding of rare SNPs contributing to the phenotype, no association was identified with common SNPs (MAF > 10%) and HDL-C in one population examined as part of the Dallas Heart Study (19), further affirming the utility and success of the "direct" candidate gene approach for this phenotype.

Although the direct approach using nsSNPs has proven successful, it is not without serious challenges. The first major challenge is the fact that cSNPs in general are fewer in number within genes compared with noncoding SNPs (noncSNPs) (42, 58). Also, cSNPs are rarer than noncSNPs, resulting in a lower average MAF among cSNPs compared with noncSNPs (38, 107). This challenge is particularly relevant to SNP discovery strategies. Currently, the goal of many large-scale SNP discovery projects is to catalogue common human DNA variation (23). Both the SeattleSNPs PGA and EGP projects resequence a sample size necessary to capture common variation (MAF > 5%) at a detection rate of 95% (55) in at least two populations: African Americans and European Americans. SNPs with a MAF < 5% would require twice the sample size (96 chromosomes) for a similar detection rate (55). Alternatively, rarer cSNPs could be more readily identified by resequencing individuals in the extreme upper and lower percentile of a specific phenotype, as did Cohen et al. (19). For both approaches, to obtain the complete catalogue of SNPs, including both coding and noncoding, will require more effort and resources.

A second challenge is that not all cSNPs are deleterious. Only approximately half of the cSNPs are nsSNPs in the general population (6, 7, 43, 84, 87, 96), with only a fraction predicted to be deleterious to protein function. Recent strides, however, have been made in the development of tools that predict in silico the effect a nsSNP has on gene function. These tools (e.g., SIFT and PolyPhen) use structural criteria and sequence homology to predict nsSNP function (15, 18, 67, 78, 102).

In our SeattleSNPs PGA and EGP data sets, 2778 (4.3%; 1 every 4 kb) of the SNPs discovered are classified as cSNPs (synonymous and nonsynonymous), of which 1396 (2.2% of all SNPs) are classified as nonsynonymous. Similar to

another large-scale resequencing survey (96), the average gene across both data sets contains four cSNPs, half of which are nsSNPs. Of all nsSNPs in the SeattleSNPs PGA and EGP, 186 are classified as "intolerant" by SIFT and "probably damaging" or "possibly damaging" by PolyPhen. Many of these potentially deleterious cSNPs do not seem to be randomly distributed across the 510-gene data set; that is, many of these cSNPs are clustered in a few genes, such as *CYP4F2* (three potentially deleterious cSNPs) in the SeattleSNPs PGA data set and *CYP2C9* (four potentially deleterious cSNPs) in the EGP data set.

A list of the potentially deleterious cSNPs was recently reviewed for the EGP data set (58). To summarize those data, 57 cSNPs out of 541 nsSNPs in the EGP were predicted to be intolerant by both SIFT and PolyPhen (58). For the SeattleSNPs PGA data set, 69 nonsynonymous cSNPs are predicted by both PolyPhen and SIFT to be deleterious to protein function (Table 1). From the 69 potentially deleterious cSNPs, 8 have a MAF $\geq$ 5% in both African-American and European-American samples; 20 have a MAF $\geq$ 5% in African-American samples, and 18 have a MAF $\geq$ 5% in European-American samples. The Pfam and Human Gene Mutation Databases were scanned in an effort to identify functionally important protein domains and previously reported phenotypic information associated with these potentially deleterious polymorphisms. The Pfam database was searched to determine if any of the 69 SeattleSNPs PGA cSNPs occur within identifiably important proteins domains. Of the 69 potentially deleterious cSNPs, 37 are located within Pfam domains (Table 1). These 37 cSNPs are good candidates for association and functional studies because they may disrupt protein folding, protein stability, or protein-protein interactions.

In addition to these 37 potentially detrimental cSNPs, of the 69 SeattleSNPs PGA nonsynomous cSNPs, 8 are associated with a known phenotype as reported in the Human Gene Mutation Database. For example, all three of the *ABO* cSNPs listed in Table 1 (P74S, R176G, and F216I) were previously identified and correspond to ABO blood group variation (69, 115). Also, *KEL* R281W was previously reported as a kell blood group variant (57), *IL4R* C431R is associated with IgE levels (41), *MC1R* R151C and R160W are associated with red hair color and vulnerability to UV-induced skin damage (34, 95), and *TNFRSF1B* M196R is associated with hyperandrogenism and polycystic ovary syndrome (72). Considering that 8 of the 69 potentially deleterious cSNPs are associated with a phenotype, these data suggest that the remaining 61 potentially deleterious cSNPs listed in Table 1 are good candidates for disease association and functional studies.

Besides the potentially deleterious nsSNPs, such as the ones described above, there are several other SNPs that should be considered candidates for direct candidate gene association studies. These include SNPs that introduce stop codons and cause premature truncation of the protein, SNPs that alter splice sites, and diallelic insertion/deletion polymorphisms that cause frame shifts. The SeattleSNPs PGA and EGP projects have identified a total of 16 SNPs that introduce a premature stop codon, 15 SNPs that alter splice sites, and 23 insertion/deletion polymorphisms that cause a frame shift with a frequency of one per 0.22 megabases resequenced, collectively (Table 2). Most of these SNPs are found only once or twice

TABLE 1 Potentially deleterious nonsynonymous cSNPs predicted by SIFT and Polyphen in the SeattleSNPs PGA data set

HUGO symbol[a]	Locus link[b]	rs (dbSNP)[c]	AA freq[d]	EA freq[e]	AA Pos.[f]	Major allele[g]	Minor allele[h]	Pfam domains[i]
ABO	28	512770	0.12	0.14	74	P	S	PF03414 Glycosyltransferase family 6
ABO	28	7853989	0.23	0	176	R	G	PF03414 Glycosyltransferase family 6
ABO	28	8176740	0.17	0.24	216	F	I	PF03414 Glycosyltransferase family 6
BDKRB2	624	2227279	0.04	0	354	G	E	
BF	629	4151667	0	0.07	9	L	H	
C2	717	4151648	0.07	0	734	R	C	PF00089 Trypsin
C3	718	—	0	0.24	314	P	L	
CKM	1158	—	0.02	0	243	G	A	PF00217 ATP:guanido phosphotransferase
CSF3R	1441	3917996	0.02	0	562	Y	H	PF00041 Fibronectin type III domain
CYP4F2	8529	3093104	0.02	0	7	S	Y	
CYP4F2	8529	3093153	0.02	0.04	185	G	V	PF00067 Cytochrome P450
CYP4F2	8529	2108622	0.09	0.17	433	V	M	PF00067 Cytochrome P450
EPHB6	2051	8177143	0.03	0	267	P	R	
F11	2160	—	0.07	0	339	C	F	PF00024 PAN domain
F12	2161	—	0.02	0	605	Y	H	PF00089 Trypsin
F13A1	2162	3024477	0	0.04	205	Y	F	
F2R	2149	2227799	0.02	0	412	S	Y	
F2RL3	9002	2227346	0.1	0	296	F	V	PF00001 7 transmembrane receptor

(*Continued*)

TABLE 1 (*Continued*)

HUGO symbol[a]	Locus link[b]	rs (dbSNP)[c]	AA freq[d]	EA freq[e]	AA Pos.[f]	Major allele[g]	Minor allele[h]	Pfam domains[i]
F5	2153	—	0.05	0	15	G	S	
F5	2153	6018	0.02	0.07	817	N	T	
F5	2153	6005	0.06	0	1146	H	Q	
F5	2153	—	0.08	0.02	1404	P	S	PF06049 Coagulation Factor V LSPD Repeat
F5	2153	—	0	0.05	2148	M	T	PF00754 F5/8 type C domain
F9	2158	4149751	0.02	0	461	T	P	
FGG	2266	6063	0	0.02	191	G	R	PF00147 Fibrinogen beta and gamma chains
HABP2	3026	7080536	0.02	0.04	534	G	E	PF00089 Trypsin
IKBKB	3551	—	0	0.02	554	R	W	
IL11RA	3590	—	0	0.02	395	R	W	
IL12B	3593	3213119	0	0.02	298	V	F	PF00041 Fibronectin type III domain
IL12RB2	3595	—	0.02	0	808	L	R	
IL17RB	55540	—	0.05	0	499	C	R	
IL21R	50615	—	0.02	0	191	R	C	
IL21R	50615	—	0.15	0	484	G	S	
IL2RA	3559	—	0.02	0	272	I	T	
IL4R	3566	1805012	0.09	0.13	431	C	R	
IL4R	3566	3024678	0.02	0.02	675	P	S	
IL8RA	3577	—	0	0.02	335	R	C	
ITGA8	8516	2298033	0.04	0.02	577	S	F	
KEL	3792	8176059	0	0.02	281	R	W	PF05649 Peptidase family M13
KLKB1	3818	3733402	0.35	0.45	143	N	S	PF00024 PAN domain

(*Continued*)

TABLE 1 (*Continued*)

HUGO symbol[a]	Locus link[b]	rs (dbSNP)[c]	AA freq[d]	EA freq[e]	AA Pos.[f]	Major allele[g]	Minor allele[h]	Pfam domains[i]
KLKB1	3818	4253379	0.02	0	358	T	A	PF00024 PAN domain
MC1R	4157	1805007	0	0.11	151	R	C	PF00001 7 transmembrane receptor
MC1R	4157	1110400	0	0.02	155	I	T	PF00001 7 transmembrane receptor
MC1R	4157	1805008	0	0.09	160	R	W	PF00001 7 transmembrane receptor
MMP9	4318	3918252	0.06	0	127	N	K	PF00413 Matrixin
NFKBIB	4793	—	0	0.02	339	R	W	
PLAT	5327	8178733	0.05	0	34	A	D	
PLAT	5327	8178747	0.02	0	136	R	S	PF00051 Kringle domain
PLAT	5327	2020921	0	0.05	164	R	W	PF00051 Kringle domain
PLAUR	5329	4251813	0.02	0	55	E	G	PF00021 u-PAR/Ly-6 domain
PLAUR	5329	4251878	0.02	0	105	R	Q	
PLAUR	5329	4760	0.02	0.12	317	L	P	
PLG	5340	4252186	0	0.02	133	H	Q	PF00051 Kringle domain
PROZ	8858	3024778	0	0.02	70	E	K	PF00594 (GLA) domain
PTGS1	5742	10306140	0.02	0.02	149	R	L	PF03098 Animal haem peroxidase
SELE	6401	3917408	0.02	0	31	M	I	PF00059 Lectin C-type domain

(*Continued*)

TABLE 1 (*Continued*)

HUGO symbol[a]	Locus link[b]	rs (dbSNP)[c]	AA freq[d]	EA freq[e]	AA Pos.[f]	Major allele[g]	Minor allele[h]	Pfam domains[i]
SELE	6401	5361	0.02	0.09	149	S	R	PF00008 EGF-like domain
SELE	6401	3917429	0.02	0	550	P	S	
SELP	6403	3917718	0.02	0	179	G	R	PF00008 EGF-like domain
SELP	6403	3917869	0	0.02	230	C	F	PF00084 Sushi domain (SCR repeat)
SERPINA5	5104	—	0.11	0.37	64	N	S	PF00079 Serpin (serine protease inhibitor)
SERPINA5	5104	—	0	0.02	115	L	P	PF00079 Serpin (serine protease inhibitor)
SERPINC1	462	—	0	0.02	30	V	E	
SFTPA1	6435	4253527	0.12	0.02	219	R	W	PF00059 Lectin C-type domain
TNFRSF1B	7133	1061622	0.18	0.2	196	M	R	
TNFRSF1B	7133	—	0.03	0	269	T	P	
TNFRSF1B	7133	—	0.02	0	301	P	R	
TRPV5	56302	4236480	0.44	0.28	154	R	H	
TYK2	7297	—	0	0.09	684	I	S	PF00069 pkinase; PF07714 Prot tyr kinase

[a]HUGO symbol.

[b]Locus Link identifier.

[c]Reference SNP cluster identifier.

[d]Estimated frequency in 24 African-American (AA) samples.

[e]Estimated frequency in 23 European-American (EA) samples.

[f]Amino acid position in coding sequence.

[g]Amino acid substitution of higher frequency allele.

[h]Amino acid substitution of lower frequency allele.

[i]Protein family.

TABLE 2 SNPs predicted to alter translation or splicing in the SeattleSNPs PGA and EGP data sets

HUGO name[a]	Type of variant	Codon[b]	Minor allele freq[c]
ABO	Frame Shift	87	0.31
ABO	Frame Shift	353	0.1
ABO	Splice		0.01
BRCA2	Frame Shift	55	0.01
BRCA2	Frame Shift	2092	0.01
BRCA2	Truncation	3326	0.01
C2	Frame Shift	281	0.01
C3AR1	Splice		0.06
CD36	Frame Shift	52	0.01
CD36	Frame Shift	334	0.01
CD36	Truncation	324	0.04
CD36	Splice		0.01
CD36	Splice		0.01
CD36	Splice		0.01
CDKL1	Frame Shift	340	0.01
CYP2C9	Frame Shift	273	0.01
EDN3	Frame Shift	189	0.01
EGF	Frame Shift	1136	0.03
ERCC4	Truncation	723	0.01
ESRRG	Splice		0.01
EXO1	Splice		0.01
GTF2H3	Frame Shift	1	0.01
HGF	Truncation	364	0.01
IL16	Truncation	572	0.19
IL17RB	Truncation	484	0.05
IL19	Splice		0.01
IL2RA	Splice		0.01
IL5RA	Frame Shift	321	0.01
IL9R	Splice		0.02
MGST1	Frame Shift	58	0.01
MGST1	Truncation	95	0.01
MGST2	Frame Shift	34	0.01
MMP19	Truncation	5	0.05

(*Continued*)

TABLE 2 (*Continued*)

HUGO name[a]	Type of variant	Codon[b]	Minor allele freq[c]
MSH6	Frame Shift	1358	0.01
MYBPC3	Truncation	70	0.01
NEIL1	Splice		0.01
ORC2L	Splice		0.01
PLA2G4C	Frame Shift	298	0.02
PLA2G4C	Frame Shift	358	0.01
POLE	Splice		0.01
POLI	Frame Shift	160	0.01
PON1	Truncation	194	0.01
PON2	Splice		0.01
RAD21	Splice		0.01
RAD23A	Truncation	11	0.01
RAD52	Truncation	346	0.03
RAD52	Truncation	415	0.05
RAG1	Frame Shift	113	0.01
SFTPA1	Truncation	242	0.01
SMUG1	Truncation	3	0.01
TAF1C	Splice		0.01
TNFAIP2	Frame Shift	105	0.04
WRN	Truncation	1406	0.01
XPA	Frame Shift	80	0.01

[a]HUGO symbol.

[b]Codon in which cSNP occurs.

[c]Minor allele frequency in either 47 SeattleSNPs PGA individuals (African Americans and European Americans combined) or 90 EGP individuals from the Polymorphism Discovery Resource Panel.

in the samples resequenced, suggesting that these SNPs may represent mutations (defined as having a MAF $< 1\%$ in the general population). A few of these SNPs are common, such as the nonsense SNP in *IL16* (Table 2), which has a MAF of 0.15 in the African-American sample and 0.22 in the European-American sample.

The final class of SNPs that could be considered for direct candidate gene association studies is the regulatory SNP group. These SNPs could include any SNP that affects regulation of gene expression without changing an amino acid of the protein. Typically, SNPs that occur in the promoter or untranslated regions (UTR) of the gene are likely candidates for SNPs that affect gene expression because it is assumed that disruption of the promoter could affect transcription factor binding or disruption of the UTR sequence could affect mRNA stability, translation, or

transportation to other components of the cell (e.g., 44). In the SeattleSNPs PGA and EGP data sets, we have identified 566 and 2060 SNPs in the 5′ and 3′ UTRs, respectively, that could be considered candidates for being regulatory SNPs.

Although the UTRs of genes are obvious candidates for regulatory SNPs, this class of SNPs can also be extended to include exonic or intronic SNPs that were once thought to be neutral polymorphisms. Shen et al. (90) demonstrated that synonymous cSNPs can alter mRNA structure, which may affect many downstream processes such as splicing, processing, and even translation. In another example, a synonymous SNP in exon 14 of the *APC* gene causes exon skipping and is associated with Familial Adenomatous Polyposis (64). The "silent" cSNP in *APC* is one of a growing list of synonymous SNPs that alter splicing (reviewed in 11). Finally, an intronic SNP in the programmed cell death 1 (*PDCD1*) gene alters the binding site of the runt-related transcription factor 1 (*RUNX1*) and is associated with systemic lupus erythematosus (77).

Seemingly, any SNP could be a regulatory SNP. The challenge, of course, is to identify the regulatory SNP among the neutral SNPs so that the regulatory SNP can be genotyped in a candidate gene association study. Currently, most computational tools concentrate on predicting whether or not a nsSNP is deleterious; however, a few tools exist that predict regulatory SNPs. Current tools have taken several approaches to mining vast amounts of experimental and sequence data available in the public domain. For example, the database TRANSFAC contains curated eukaryotic transcription factor and DNA binding specificities from the literature (114), and the Eukaryotic Promoter Database (EPD) contains eukaryotic polymerase II promoters experimentally derived from transcriptional start sites (86). Many algorithms that predict transcriptional binding sites rely on the observation that transcriptional control is generally conserved across species. These algorithms [ConSite (85), CORG (30), PromoLign (117), rVISTA (60), and TraFaC (48)] use human-mouse orthologous sequence alignments as well as other cross-species alignments to identify putative transcription factor binding sites and other *cis*-regulatory elements. Other tools, such as the web-based PupaSNP (24) and ESEfinder (12), include the prediction of exonic splicing enhancers as potential functional SNPs.

Approximately 50% of the human genomic sequence is repetitive, and only 5% is predicted to be coding (46). Nearly half of the human genome has no known function. Recognizing the need to further annotate the list of functional SNPs, the National Human Genome Research Institute at the National Institutes of Health recently launched a collaborative project entitled the "Encyclopedia of DNA Elements" or ENCODE (21). The goal of ENCODE is to identify all the functional elements of the human genome. To do this, diverse yet complementary computational and experimental approaches will be developed in a high-throughput manner. Already, computational comparisons of genome sequences across species, such as the mouse/human comparison where 40% of the human sequence aligns to the mouse sequence (66), have identified interesting conserved noncoding regions that are candidates for being regulatory regions in the human genome (reviewed in 63, 71, 94). Ideally, ENCODE would also include high-throughput experimental

approaches that would confirm functional elements identified through computational approaches. Traditionally, these experimental approaches include DNA footprinting, gel shift assays, deletion constructs, DNASE I hypersensitivity studies, and gene trapping (including promoter and enhancer trapping) (reviewed in 31, 71). All of these approaches are time intensive at the single gene level and will require further attention to make them amenable for high-throughput identification of functional DNA elements in the human genome.

ASSOCIATION STUDIES AND SINGLE NUCLEOTIDE POLYMORPHISMS: THE "INDIRECT" APPROACH

The "indirect" approach to genetic association studies differs from the direct approach described above in that the causal SNP is not assayed directly. The indirect approach is much like a linkage study in that the study design assays many presumably neutral markers and makes no assumption on the location of the causative gene or locus. Linkage studies are family based and rely on recombination events within the pedigree to narrow the genomic region that contains the causative gene segregating within the families being studied. The indirect genetic association study is most often a case-control study drawn from the general population. Although this review focuses on indirect candidate gene association studies, this approach can be used to interrogate whole genomes and genomic regions (reviewed in 8). Like a linkage study, the indirect approach also relies on recombination to narrow the genomic region related to the phenotype. The difference is that an association study, because it is drawn from a population, uses a measure of allelic association or site correlation, known as linkage disequilibrium (LD), to detect historical recombination. The assumption is that the assayed or genotyped SNPs will be in LD or associated with the causative SNP; thus, the assayed SNP would be overrepresented among cases compared with controls because it is highly correlated with the disease-causing SNP.

The success of an indirect association study, whether at the whole genome or candidate gene level, hinges on several assumptions and parameters. The first assumption is that the disease or phenotype in question has a strong (or measurable) genetic component. The average gene in our SeattleSNPs PGA and EGP data set contains approximately 126 SNPs, of which most are neutral polymorphisms. Therefore, the presence of genetic variation alone (or even extremely high or low levels of variation) does not provide evidence that a phenotype associated with the candidate gene has a strong genetic component. The evidence for a strong genetic component is usually derived from twin studies and family studies. Although these studies are popular and important sources of evidence, it is important to realize that heritability estimates depend on the population being studied as well as the time and location of the study (reviewed in 65, 109). Also, it is often incorrectly assumed that strong heritability indicates that there is a single major gene underlying the disease or trait. Finally, most studies estimating heritability are biased in that

they overestimate the effects of major genes. These limitations on our ability to estimate the strength of the genetic component could have a significant impact on our ability to design the optimal study to detect disease-susceptibility loci for a given phenotype.

A second assumption commonly made at the onset of study design is that, for a given common human disease, only a few common variants are associated with it (14, 22, 56). A few examples exist in which "common" variants are consistently associated with a common disease phenotype [e.g., *CARD15* and Crohn's disease (45); *APOE* and Alzheimer's disease (33)]. However, some investigators argue that not all common human disease can be attributed to a few common variants, and that it is more likely that several rare variants at several sites (genetic and allelic heterogeneity) could result in the phenotypes being studied (74, 75, 112). To further complicate matters, the notion of "common" is arbitrary, with some investigators defining this as an allele with >20% MAF and others defining this as an allele with >1% MAF. This seemingly arbitrary decision to limit the set of potential SNPs (usually on the basis of genotyping costs and effort) to those with a certain allele frequency may have serious consequences on the power of the study to detect a difference between cases and controls. Recent work has established that the most efficient and powerful studies that can detect disease-susceptibility alleles are those in which the allele frequency differences between the genotyped SNP and the disease-causing or -susceptibility SNP are small (reviewed in 119). If the difference in allele frequencies between the genotyped SNP and the disease-causing SNP is large, especially in the case where the disease-causing SNP is rarer than the genotyped SNP (51), it is likely that only studies searching for genetic determinants of large phenotypic effects will be successful (119).

A third assumption made in designing indirect association studies is that there will be useful levels of LD within the genomic region of the population studied and that these levels of LD can help determine which and how many markers should be genotyped in the study. In general, regions of the genome with large stretches of LD are desirable in an association study because fewer markers have to be genotyped; however, these regions become less desirable when the investigator tries to tease the disease-causing SNPs from the other SNPs in LD with it (76, 80). Despite the double-edged sword effect of LD, there is great interest in describing LD properties across the genome and developing strategies to choose SNPs for genotyping based on these patterns.

Using the resequencing data generated across the 510 SeattleSNPs PGA and EGP data sets, we can describe LD across candidate genes using the measure r^2 (Figure 2). Pair-wise LD can also be measured using D', which is useful in describing historical recombination in a sample. The r^2 measure is preferred in this context because there is an inverse relationship between this measure and the power to detect an association (reviewed in 76, 110). As demonstrated in Figure 2*a*,*b*,*c*, the strength of LD can vary dramatically across candidate genes. For example, for

two similar-sized genes, *TGFB3* (~24 kb) and *IL1R2* (~23 kb), the strength and extent of LD is different even within the same European-American sample (Figure 1*a*,*c*). The strength and extent of LD is also different within the same gene but across population samples (Figure 2*a*,*b*). The difference in LD patterns between populations can be an asset in specific situations where the phenotype is the same between the two populations, but the levels of LD are different so that an association can be detected with a minimal number of markers, but the causative SNP could also be identified (61, 118). This is a strategy that essentially circumvents the double-edged sword effect often observed with studying genotype-phenotype correlations in only one population.

The strength and extent of LD is influenced by several factors. It is well documented that LD decays with increasing physical distance (1, 3). Also, African-descent populations typically have less LD or shorter-ranged LD compared with European-descent populations due to differences in population history (35, 38, 53, 79, 84, 87, 93, 106). This is also evident in our SeattleSNPs PGA candidate gene set. Other population structures, such as isolated populations (54) and recently admixed populations (13), are predicted to have long stretches of LD compared with other populations, although there is evidence that contradicts these predictions (32, 59).

Another factor that influences the structure of LD is natural selection. Evidence for natural selection can be assessed in sequence data using several statistics that summarize the allele frequency distribution within samples for the region of interest. One such popular test statistic is Tajima's D, which tests departure from neutrality using two measures of nucleotide diversity (103). Figure 3 plots the values of Tajima's D for 180 SeattleSNPs genes calculated for the African-American and European-American samples. The average Tajima's D is –0.51 in the African-American sample and 0.18 in the European-American sample. For the European-American sample, a few genes have Tajima's D values >2 or ≤ 2 (Figure 3), values that may be considered extreme for this test statistic. A large negative Tajima's D indicates that the gene has an excess of rare variants, suggesting the gene has experienced positive selection. A large positive Tajima's D indicates the gene has an excess of intermediate-frequency alleles, suggesting the gene has experienced balancing selection. Both these extremes in Tajima's D, however, can be explained by population demography or history; therefore, a more careful examination of these values must be performed before conclusions about departures from neutrality can be made for these genes (2).

Finally, another major factor influencing the structure of LD is recombination. The relationship between the strength of LD and the amount of recombination is evident at both the candidate gene level (e.g., 17) and the genome-wide level. For example, in a detailed study of chromosome 22 in European-American samples (ascertained from available Center d'Etude du Polymorphisme Humain samples), Dawson et al. (29) demonstrated that LD tends to be strongest in areas of little or no recombination, in addition to decaying as the physical distance increases. Also,

Clark et al. (16) examined 4833 SNPs typed in 538 clusters across the human genome and identified a negative correlation between the LD metric r^2 and the recombination rate ρ.

Collectively, all of these factors (population history, natural selection, and recombination) make it difficult to predict a priori which regions of the human genome will have levels of LD acceptable for genetic association studies. Despite this limitation, several groups are optimistic that the human genome can be described based on patterns of LD. Both candidate genes (101) and regions in the human genome (27) have been described as having regions of high LD (termed "blocks") separated or punctuated by regions of low LD, presumably caused by hot spots of recombination. The "haplotype blocks" contain a few common haplotypes accounting for most of the chromosomes assayed (27). Gabriel et al. (36) formalized the definition of haplotype blocks and demonstrated that haplotype blocks existed across the human genome in several different populations, although African-descent samples had a greater number of short blocks compared with non-African-descent samples.

Studies on haplotype block structure (36, 70) were immediately influential because they provided the framework from which to choose SNPs for genotyping in a genetic association study (49). These data also provided impetus for the International HapMap Project, an international collaborative effort designed to describe LD across the human genome in several populations and the tagSNPs needed to capture genetic diversity that can be applied in future genetic association studies (105). However, since the publication of these findings, several investigators have demonstrated that, even though hot spots of recombination may be a common feature of the human genome (25, 62), haplotype blocks can exist without their presence (73, 111). Furthermore, studies show that the number and size of haplotype blocks depend on marker density (50, 98), the MAF cutoff imposed on the data set (88, 100), and block definition (89). These recent findings will no doubt increase the usefulness of the International HapMap Project as researchers investigate new avenues to describe LD and genetic diversity across the human genome and ways to apply this knowledge in well-designed genetic association studies.

As the genetics community waits for the whole-genome association study resources to be developed by the HapMap, many investigators are continuing work on candidate genes in association studies. Several useful algorithms have been developed to choose SNPs for genotyping that can be applied to candidate genes. Most algorithms available today are based on choosing SNPs that represent common haplotypes (usually defined as having a frequency of $>5\%$ in the sample) or maximize haplotype diversity within blocks (49, 99, 116). The disadvantage of this approach is that these algorithms require haplotypes. To date, most haplotype data represent inferred haplotypes, not molecularly determined haplotypes. It is clear that genomic regions with high haplotype diversity are difficult to phase correctly (68) and that, no matter the algorithm used, a fraction of the inferred

haplotypes is incorrect (97). Also, we recently showed in the SeattleSNPs PGA data set that the number of inferred haplotypes per gene varies substantially across genes and between populations (26) (Figure 4). For example, the average number of inferred haplotypes per gene is 34 for the SeattleSNPs PGA and 38 for the EGP (Figure 4*a*). In contrast, the range for the number of haplotypes per gene is quite broad: 7 to 94 in the SeattleSNPs PGA and 2 to 175 in the EGP (Figure 4*a*). Also, as shown previously (4, 26), the average number of haplotypes per gene, as well as the range of haplotypes per gene, is greater in the African-American sample compared with the European-American sample: 25 versus 15 (Figure 4*b*). In general, the larger genes have the most haplotypes because there is a positive correlation between the increasing number of SNPs and the increasing number of haplotypes. However, there are always exceptions, such as *CCND2*, a gene with 100 inferred haplotypes that is only 33.5 kb in size and contains 48 SNPs with MAF $\geq$ 5%. We found that haplotype diversity was high in a few genes such that no common haplotypes (>5% frequency) were inferred for that sample (26). These data demonstrate that a proportion of genes have high haplotype diversity, perhaps making them less amenable to haplotype tagging approaches.

One tagSNP selection algorithm that does not require the inference of haplotypes across the entire gene is LDSelect, an algorithm that groups correlated SNPs from genotype data using linkage disequilibrium (10). At an empirically determined r^2 threshold of 0.64 and a minor allele frequency $\geq$5%, we identified 3709 and 1807 tagSNPs for genotyping in the African-American and European-American samples, respectively, across 180 genes in the SeattleSNPs PGA data set. The average number of tagSNPs per gene is 21 for the African-American sample and 10 for the European-American sample. Even though twice as many tagSNPs must be genotyped on average for the African-American sample compared with the European-American sample at this particular r^2 threshold, the number of SNPs required for genotyping represents a great savings because the tagSNPs only account for 36% of the total number of SNPs with MAF $\geq$ 5% in the African-American sample.

One important aspect of any tagSNP algorithm is the fact that a set of tagSNPs is population specific. That is, tagSNPs determined for a European-descent sample should not be applied to an African-descent population. Also, tagSNPs should not be determined in a stratified or mixed population (such as the Polymorphism Discovery Resource Panel) and then applied to a defined population. There is evidence, however, that tagSNPs can be chosen in one sample and applied in another sample of similar race/ethnicity (51). The difference across tagSNPs sets for different populations stems, in part, from differences in population history. As discussed above, it is well known that LD patterns differ across populations. The number of SNPs is different between two populations, with African-descent populations typically having the most variants compared with other populations. Also, a previous analysis of the SeattleSNPs PGA data set demonstrated that

minor allele frequencies for the same sites can differ dramatically across different populations (9). Thus, a common SNP in one sample may not necessarily be a common SNP in a second sample that differs by race/ethnicity.

VASCULAR CELL ADHESION MOLECULE 1 (*VCAM1*): AN AVERAGE GENE IN AN AVERAGE STUDY

This review has described two basic approaches for candidate gene association studies: direct and indirect. As a working example of these approaches, we apply the concepts described above to a SeattleSNPs PGA gene, vascular cell adhesion molecule 1 (*VCAM1*). *VCAM1* is a member of the immuoglobulin gene superfamily induced by cytokines, and its upregulation has been noted at atherosclerotic lesions in mice (28). *VCAM1* is located at 1p32-p31 and is approximately 22.8 kb in size. Resequencing *VCAM1* in 47 individuals revealed 113 unique SNPs, 102 of which are present in the African-American sample and 39 of which are present in the European-American sample. The nucleotide diversity (π) of *VCAM1* is 6.55×10^{-4} and 3.93×10^{-4} for the African-American and European-American sample, respectively, and Tajima's D is negative in the African-American sample and positive in the European-American sample (-1.26×10^{-4} and 0.03×10^{-4}, respectively). Of the 113 total SNPs, 43 have a MAF $\geq$ 5% (48 in the African-American sample and 23 in the European-American sample).

For a direct association study approach, we are interested in genotyping functional SNPs within *VCAM1*. *VCAM1* contains six SNPs in the coding region. For the six cSNPs, two are synonymous and four are nonsynonymous. One of the nsSNPs, S318F (African-American MAF = 0.02; European-American MAF = 0.00), is predicted by PolyPhen to be possibly damaging/intolerant, making it a candidate for a direct association study. In examining the promoter region of *VCAM1*, we find a total of 10 SNPs within 2 kb of the start of transcription. We used Alibaba2.1 (see Electronic Databases) to identify SNPs occurring within predicted transcription factor binding sites (39). Among these promoter SNPs, four occur within predicted transcription factor binding sites. Specifically, an T/C SNP –833 bp, T/G SNP –1599 bp, C/T SNP –2021 bp, and A/C SNP –2062 bp from the start of translation are predicted to fall within a Oct-1 or HNF-3 or C/EBP alpha, AP-1, Oct-1, and an ICSBP transcription factor binding site, respectively. Therefore, these SNPs should also be considered for an association study because it is possible that they have a functional effect by influencing the binding of transcription factors to the promoter region, resulting in allele-specific levels of *VCAM1* gene expression.

For an indirect association study approach, we are interested in genotyping SNPs that are either the causative SNP or the SNP in LD with the causative SNP. Using LDSelect (10) with an r^2 threshold of 0.64 and a MAF threshold of $\geq$5%, we would choose 30 tagSNPs for the African-American sample and

11 tagSNPs in the European-American sample for genotyping in *VCAM1*. Alternatively, we could employ a haplotype tagging approach to choose tagSNPs. At a MAF $\geq$ 5%, the African-American sample has 44 inferred haplotypes and the European-American sample has 22 inferred haplotypes (PHASEv2.1; see Electronic Databases). Using the default settings for the D′ block definition in HaploBlockFinder (116), 29 tagSNPs in the African-American sample and 7 tagSNPs in the European-American sample were identified from inferred haplotypes in *VCAM1* for genotyping.

CONCLUSIONS

We demonstrate here that much has been uncovered about the patterns of human gene variation through recent DNA variation discovery efforts. However, much remains to be learned about how these patterns can be applied to human genetic association studies to identify the loci involved in common human diseases. The data presented here suggest that specific approaches and methods may be appropriate for a proportion of the genes in the human genome, but may not be powerful or appropriate for other genes. This realization that one method does not fit all genetic association studies should stimulate creative thinking for alternative approaches to applying DNA variation in association studies so that the successes of the Human Genome Project will translate into successes of genomic medicine and public health for all populations.

ACKNOWLEDGMENTS

We thank members of the SeattleSNPs PGA team (M. Ahearn, T. Armel, C. Bertucci, D. Carrington, L. Daniels, S. Da Ponte, M. Eberle, N. Hastings, E. Johanson, P. Keyes, L. Kruglyak, S. Kuldanek, N. Rajkumar, M. Rieder, T. Shaffer, E. Toth, M. Wong, and Q. Yi), members of the EGP team (B. Borrayo, C. Cassidy, S. Chambers, M-W Chung, M. Daniels, T. Downing, H. Gildersleeve, T. Jackson, E. Johnson, B. Leithauser, R. Livingston, I. McFarland, K. Miyamoto, M. Montoya, C. Nguyen, D. Nguyen, A. Olson, C. Park, C. Poel, P. Robertson, W. Schackwitz, A. Sherwood, K. Sherwood, J. Swanson, E. Torskey, L. Witrak, and B. Yool), and other members of the Nickerson laboratory (C. Baier, T. Bhangale, M. Bloomfield, E. Calhoun, C. Carlson, B. Howie, P. Lee, R. Mackelprang, C. Shephard, J. Sloan, J. Smith, Z. Stednick, and M. Wimpee) for generating the high-quality DNA variation data presented in this manuscript. We also thank J. Calhoun for providing the dbSNP data. This work was supported by grants from the National Heart, Lung, and Blood Institute (Program for Genomic Applications, HL66682–05; Pharmacogenetics Network for Cardiovascular Risk Therapy, HL069757) and the National Institute of Environmental Health Sciences (Environmental Genome Project, ES15478).

ELECTRONIC DATABASES

Alibaba2.1
http://darwin.nmsu.edu/~molb470/fall2003/Projects/solorz/aliBaba_2_1.htm

dbSNP
http://www.ncbi.nlm.nih.gov/SNP/index.html

Environmental Genome Project
http://egp.gs.washington.edu

Genbank
http://www.ncbi.nlm.nih.gov/Genbank/index.html

HaploBlockFinder
http://cgi.uc.edu/cgi-bin/kzhang/haploBlockFinder.cgi

Human Gene Mutation Database
http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html

LocusLink
http://www.ncbi.nlm.nih.gov/projects/LocusLink/

PHASEv2.1
http://www.stat.washington.edu/stephens/software.html

Pfam
http://www.sanger.ac.uk/Software/Pfam

SeattleSNPs Program for Genomic Applications
http://pga.gs.washington.edu

Visual Genotype 2.0 (VG2)
http://pga.gs.washington.edu/VG2.html

The *Annual Review of Genomics and Human Genetics* is online at http://genom.annualreviews.org

LITERATURE CITED

1. Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, et al. 2001. Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* 68:191–97
2. Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, et al. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* 2:e286
3. Bhangale TR, Rieder MJ, Livingston R, Nickerson DA. 2005. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.* 14:59–69
4. Bonnen PE, Wang PJ, Kimmel M,

Chakraborty R, Nelson DL. 2002. Haplotype and linkage disequilibrium architecture for human cancer-associated genes. *Genome Res.* 12:1846–53

5. Botstein D, Risch N. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* 33(Suppl.):228–37
6. Cambien F, Poirier O, Nicaud V, Herrmann SM, Mallet C, et al. 1999. Sequence diversity in 36 candidate genes for cardiovascular disorders. *Am. J. Hum. Genet.* 65:183–91
7. 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.* 22:231–38
8. Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. 2004. Mapping complex disease loci in whole-genome association studies. *Nature* 429:446–52
9. Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, et al. 2003. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat. Genet.* 33:518–21
10. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, et al. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* 74:106–20
11. Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.* 3:285–98
12. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. 2003. ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.* 31:3568–71
13. Chakraborty R, Weiss KM. 1988. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl. Acad. Sci. USA* 85:9119–23
14. Chakravarti A. 1999. Population genetics–making sense out of sequence. *Nat. Genet.* 21:56–60
15. Chasman D, Adams RM. 2001. Predicting the functional consequences of nonsynonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* 307: 683–706
16. Clark AG, Nielsen R, Signorovitch J, Matise TC, Glanowski S, et al. 2003. Linkage disequilibrium and inference of ancestral recombination in 538 single-nucleotide polymorphism clusters across the human genome. *Am. J. Hum. Genet.* 73:285–300
17. Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, et al. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* 63:595–612
18. Clifford RJ, Edmonson MN, Nguyen C, Buetow KH. 2004. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics* 20:1006–14
19. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, et al. 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 305:869–72
20. Collins FS, Brooks LD, Chakravarti A. 1998. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* 8:1229–31
21. Collins FS, Green ED, Guttmacher AE, Guyer MS. 2003. A vision for the future of genomics research. *Nature* 422:835–47
22. Collins FS, Guyer MS, Charkravarti A. 1997. Variations on a theme: cataloging human DNA sequence variation. *Science* 278:1580–81
23. Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, et al. 1998. New goals for the U.S. Human Genome Project: 1998–2003. *Science* 282:682–89
24. Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, et al. 2004.

PupaSNP Finder: a web tool for finding SNPs with putative effect at transcription level. *Nucleic Acids Res.* 32:W242–48

25. Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, et al. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat. Genet.* 36:700–6
26. Crawford DC, Carlson CS, Rieder MJ, Carrington DP, Yi Q, et al. 2004. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am. J. Hum. Genet.* 74:610–22
27. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. 2001. High-resolution haplotype structure in the human genome. *Nat. Genet.* 29:229–32
28. Dansky HM, Barlow CB, Lominska C, Sikes JL, Kao C, et al. 2001. Adhesion of monocytes to arterial endothelium and initiation of atherosclerosis are critically dependent on vascular cell adhesion molecule-1 gene dosage. *Arterioscler. Thromb. Vasc. Biol.* 21:1662–67
29. Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, et al. 2002. A first-generation linkage disequilibrium map of human chromosome 22. *Nature* 418:544–48
30. Dieterich C, Wang H, Rateitschak K, Luz H, Vingron M. 2003. CORG: a database for COmpartive Regulatory Genomics. *Nucleic Acids Res.* 31:55–57
31. Durick K, Mendlein J, Xanthopoulos KG. 1999. Hunting with traps: genome-wide strategies for gene discovery and functional analysis. *Genome Res.* 9:1019–25
32. Eaves IA, Merriman TR, Barber RA, Nutland S, Tuomilehto-Wolf E, et al. 2000. The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 25:320–23
33. Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, et al. 1997. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA* 278:1349–56
34. Frandberg PA, Doufexis M, Kapas S, Chhajlani V. 1998. Human pigmentation phenotype: a point mutation generates nonfunctional MSH receptor. *Biochem. Biophys. Res. Commun.* 245:490–92
35. Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, et al. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.* 69:831–43
36. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, et al. 2002. The structure of haplotype blocks in the human genome. *Science* 296:2225–29
37. Glazier AM, Nadeau JH, Aitman TJ. 2002. Finding genes that underlie complex traits. *Science* 298:2345–49
38. Goddard KA, Hopkins PJ, Hall JM, Witte JS. 2000. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.* 66:216–34
39. Grabe N. 2002. AliBaba2: context specific identification of transcription factor binding sites. *In Silico Biol.* 2:S1–S15
40. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, et al. 1983. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 306:234–38
41. Hackstein H, Hofmann H, Bohnert A, Bein G. 1999. Definition of human interleukin-4 receptor alpha chain haplotypes and allelic association with atopy markers. *Hum. Immunol.* 60:1119–27
42. Haga H, Yamada R, Ohnishi Y, Nakamura Y, Tanaka T. 2002. Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190,562 genetic variations in the human genome. Single-nucleotide polymorphism. *J. Hum. Genet.* 47:605–10

43. 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.* 22:239–47
44. Huang JL, Gao PS, Mathias RA, Yao TC, Chen LC, et al. 2004. Sequence variants of the gene encoding chemoattractant receptor expressed on Th2 cells (CRTH2) are associated with asthma and differentially influence mRNA stability. *Hum. Mol. Genet.* 13:2691–97
45. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, et al. 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 411:599–603
46. International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:806–921
47. International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931–45
48. Jegga AG, Sherwood SP, Carman JW, Pinski AT, Phillips JL, et al. 2002. Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res.* 12:1408–17
49. Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, et al. 2001. Haplotype tagging for the identification of common disease genes. *Nat. Genet.* 29:233–37
50. Ke X, Cardon LR. 2003. Efficient selective screening of haplotype tag SNPs. *Bioinformatics* 19:287–88
51. Ke X, Durrant C, Morris AP, Hunt S, Bentley DR, et al. 2004. Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Hum. Mol. Genet.* 13:2557–65
52. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, et al. 1989. Identification of the cystic fibrosis gene: genetic analysis. *Science* 245:1073–80
53. Kidd JR, Pakstis AJ, Zhao H, Lu RB, Okonofua FE, et al. 2000. Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. *Am. J. Hum. Genet.* 66:1882–99
54. Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 22: 139–44
55. Kruglyak L, Nickerson DA. 2001. Variation is the spice of life. *Nat. Genet.* 27: 234–36
56. Lander ES. 1996. The new genomics: global views of biology. *Science* 274:536–39
57. Lee S, Wu X, Son S, Naime D, Reid M, et al. 1996. Point mutations characterize KEL10, the KEL3, KEL4, and KEL21 alleles, and the KEL17 and KEL11 alleles. *Transfusion* 36:490–94
58. Livingston RJ, von Niederhausern A, Jegga A, Crawford DC, Carlson CS, et al. 2004. Pattern of sequence variation across 213 environmental response genes. *Genome Res.* 14:1821–31
59. Lonjou C, Collins A, Morton NE. 1999. Allelic association between marker loci. *Proc. Natl. Acad. Sci. USA* 96:1621–26
60. Loots GG, Ovcharenko I. 2004. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.* 32:W217–21
61. McKenzie CA, Abecasis GR, Keavney B, Forrester T, Ratcliffe PJ, et al. 2001. Trans-ethnic fine mapping of a quantitative trait locus for circulating angiotensin I-converting enzyme (ACE). *Hum. Mol. Genet.* 10:1077–84
62. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, et al. 2004. The fine scale structure of recombination rate variation in the human genome. *Science* 304:581–84
63. Miller W, Makova KD, Nekrutenko A, Hardison RC. 2004. Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* 5:15–56

64. Montera M, Piaggio F, Marchese C, Gismondi V, Stella A, et al. 2001. A silent mutation in exon 14 of the APC gene is associated with exon skipping in a FAP family. *J. Med. Genet.* 38:863–67
65. Mountain JL, Risch N. 2004. Assessing genetic contributions to phenotypic differences among 'racial' and 'ethnic' groups. *Nat. Genet.* 36:S53
66. Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–62
67. Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31:3812–14
68. Niu T. 2004. Algorithms for inferring haplotypes. *Genet. Epidemiol.* 27:334–47
69. Ogasawara K, Yabe R, Uchikawa M, Saitou N, Bannai M, et al. 1996. Molecular genetic analysis of variant phenotypes of the ABO blood group system. *Blood* 88:2732–37
70. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, et al. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294:1719–23
71. Pennacchio LA, Rubin EM. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* 2: 100–9
72. Peral B, San Millan JL, Castello R, Moghetti P, Escobar-Morreale HF. 2002. Comment: the methionine 196 arginine polymorphism in exon 6 of the TNF receptor 2 gene (TNFRSF1B) is associated with the polycystic ovary syndrome and hyperandrogenism. *J. Clin. Endocrinol. Metab.* 87:3977–83
73. Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, et al. 2003. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat. Genet.* 33:382–87
74. Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex disease? *Am. J. Hum. Genet.* 69:124–37
75. Pritchard JK, Cox NJ. 2002. The allelic architecture of human disease genes: common disease-common variant. . .or not? *Hum. Mol. Genet.* 11:2417–23
76. Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69:1–14
77. Prokunina L, Castillejo-Lopez C, Oberg F, Gunnarsson I, Berg L, et al. 2002. A regulatory polymorphism in PDCD1 is associated with susceptibility to systemic lupus erythematosus in humans. *Nat. Genet.* 32:666–69
78. Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30:3894–900
79. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, et al. 2001. Linkage disequilibrium in the human genome. *Nature* 411:199–204
80. Rieder MJ, Taylor SL, Clark AG, Nickerson DA. 1999. Sequence variation in the human angiotensin converting enzyme. *Nat. Genet.* 22:59–62
81. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, et al. 1989. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* 245:1066–73
82. Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* 273:1516–17
83. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–33
84. Salisbury B, Pungliya M, Choi JY, Jiang R, Sun JX, et al. 2003. SNP and haplotype variation in the human genome. *Mutat. Res.* 526:53–61
85. Sandelin A, Wasserman WW, Lenhard B. 2004. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.* 32:W249–52
86. Schmid CD, Praz V, Delorenzi M, Perier

R, Bucher P. 2004. The eukaryotic promoter database EPD: the impact of in silico primer extension. *Nucleic Acids Res.* 32:D82–85

87. Schneider JA, Pungliya MS, Choi JY, Jiang R, Sun XJ, et al. 2003. DNA variability of human genes. *Mech. Ageing Dev.* 124:17–25
88. Schulze TG, Zhang K, Chen YS, Akula N, Sun F, et al. 2004. Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. *Hum. Mol. Genet.* 13:335–42
89. Schwartz R, Halldorsson BV, Bafna V, Clark AG, Istrail S. 2003. Robustness of inference of haplotype block structure. *J. Comput. Biol.* 10:13–19
90. Shen LX, Basilion JP, Stanton VP Jr. 1999. Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc. Natl. Acad. Sci. USA* 96: 7871–76
91. Sherry ST, Ward M, Sirotkin K. 1999. dbSNP–database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* 9:677–79
92. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29:308–11
93. Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A. 2003. Linkage disequilibrium patterns of the human genome across populations. *Hum. Mol. Genet.* 12:771–76
94. Sidow A. 2002. Sequence first. Ask questions later. *Cell* 111:13–16
95. Smith R, Healy E, Siddiqui S, Flanagan N, Steijlen PM, et al. 1998. Melanocortin 1 receptor variants in an Irish population. *J. Invest. Dermatol.* 111:119–22
96. Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, et al. 2001. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 293:489–93
97. Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* 73:1162–69
98. Stram DO. 2004. Tag SNP selection for association studies. *Genet. Epidemiol.* 27: 365–74
99. Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, et al. 2003. Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the multiethnic cohort study. *Hum. Hered.* 55:27–36
100. Stumpf MPH. 2004. Haplotype diversity and SNP frequency dependence in the description of genetic variation. *Eur. J. Hum. Genet.* 12:469–77
101. Subrahmanyan L, Eberle MA, Clark AG, Kruglyak L, Nickerson DA. 2001. Sequence variation and linkage disequilibrium in the human T-cell receptor beta (TCRB) locus. *Am. J. Hum. Genet.* 69: 381–95
102. Sunyaev S, Ramensky V, Koch I, Lathe W III, Kondrashov AS, et al. 2001. Prediction of deleterious human alleles. *Hum. Mol. Genet.* 10:591–97
103. Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–95
104. The Huntington's Disease Collaborative Research Group. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* 72:971–83
105. The International HapMap Consortium. 2003. The International HapMap Project. *Nature* 426:789–96
106. Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, et al. 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* 271:1380–87
107. Tsunoda T, Lathrop GM, Sekine A, Yamada R, Takahashi A, et al. 2004. Variation of gene-based SNPs and linkage disequilibrium patterns in the human genome. *Hum. Mol. Genet.* 13:1623–32

108. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291:1304–51
109. Vitzthum VJ. 2003. A number no greater than the sum of its parts: the use and abuse of heritability. *Hum. Biol.* 75:539–58
110. Wall JD, Pritchard JK. 2003. Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet.* 4:587–97
111. Wang N, Akey JM, Zhang K, Chakraborty R, Jin L. 2002. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.* 71:1227–34
112. Weiss KM, Terwilliger JD. 2000. How many diseases does it take to map a gene with SNPs? *Nat. Genet.* 26:151–57
113. Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, et al. 2004. Database resources for the National Center for Biotechnology Information: update. *Nucleic Acids Res.* 32:D35–40
114. Wingender E, Chen X, Hehl R, Karas H, Liebich I, et al. 2001. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* 28:316–19
115. Yamamoto F, Clausen H, White T, Marken J, Hakomori S. 1990. Molecular genetic basis of the histo-blood group ABO system. *Nature* 345:229–33
116. Zhang K, Jin L. 2003. HaploBlockFinder: haplotype block analyses. *Bioinformatics* 19:1300–1
117. Zhao T, Chang L-W, McLeod HL, Stormo GD. 2004. PromoLign: a database for upstream region analysis and SNPs. *Hum. Mutat.* 23:534–39
118. Zhu X, McKenzie CA, Forrester T, Nickerson DA, Broeckel U, et al. 2000. Localization of a small genomic region associated with elevated ACE. *Am. J. Hum. Genet.* 67:1144–53
119. Zondervan KT, Cardon LR. 2004. The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* 5:89–100

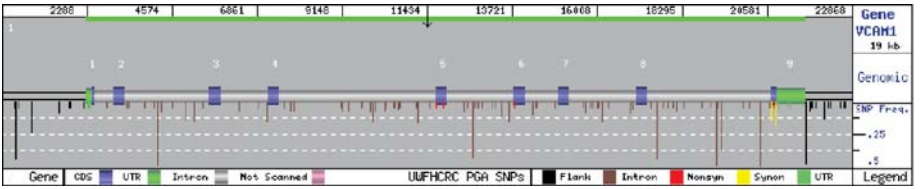

Figure 1 GeneSNPs view of VCAM1. VCAM1, a typical gene resequenced by the SeattleSNPs PGA, is located on 1p32-p31 and is ~22.8 kb long. VCAM1 was resequenced in 47 individuals and has 113 single nucleotide polymorphisms (SNPs). Six of the SNPs are coding SNPs, of which two are synonymous and four are nonsynonymous. The GeneSNPs view of each gene, including VCAM1 in this figure, displays exons (*purple boxes*), introns (*gray*), and untranslated regions (*green*) by color. The vertical lines represent SNPs, and the lengths of the vertical lines represent the minor allele frequency for the SNP. Each SNP is color coded to represent the location of the SNP within the gene: flanking sequence (*black*), intron (*brown*), exon (*red* for nonsynonymous and *yellow* for synonymous), and untranslated region (*green*).

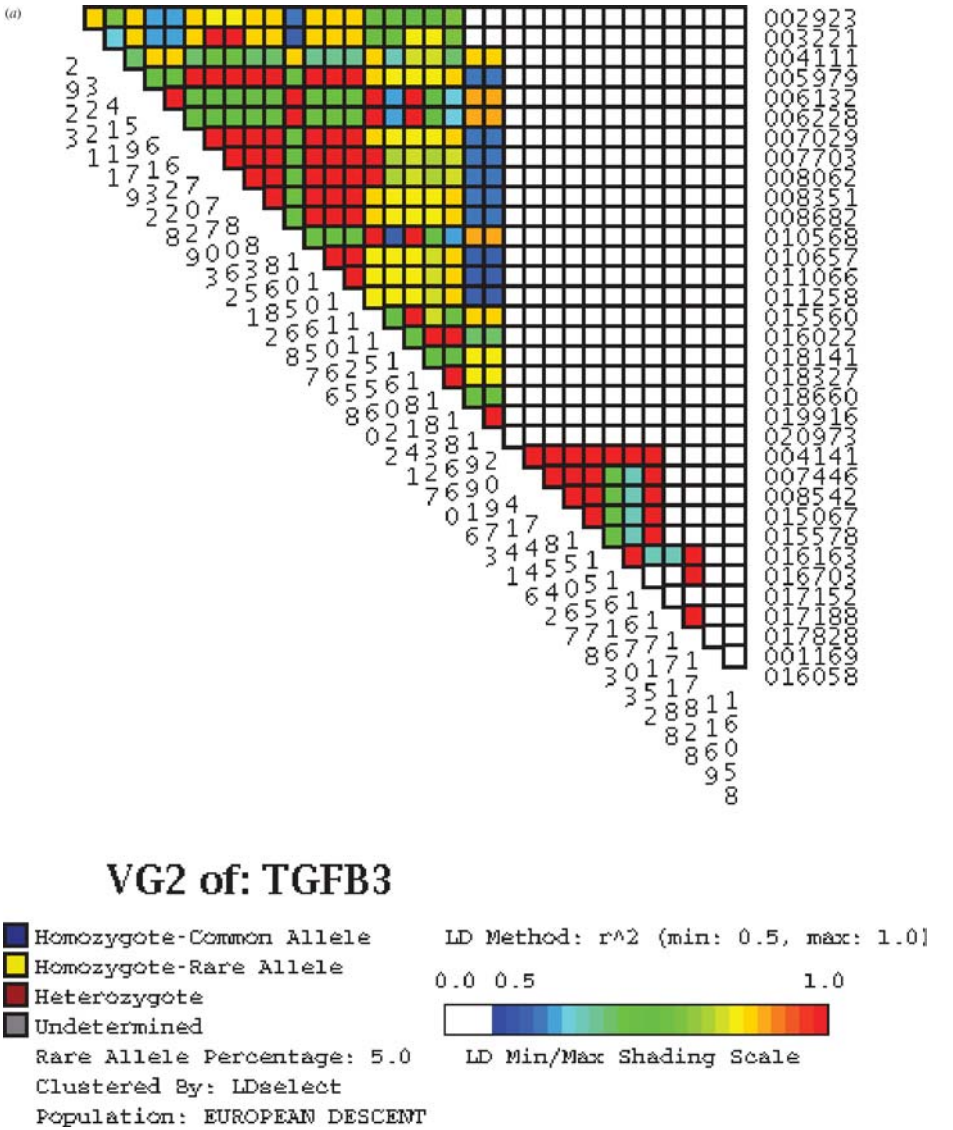

Figure 2 (*Continued*)

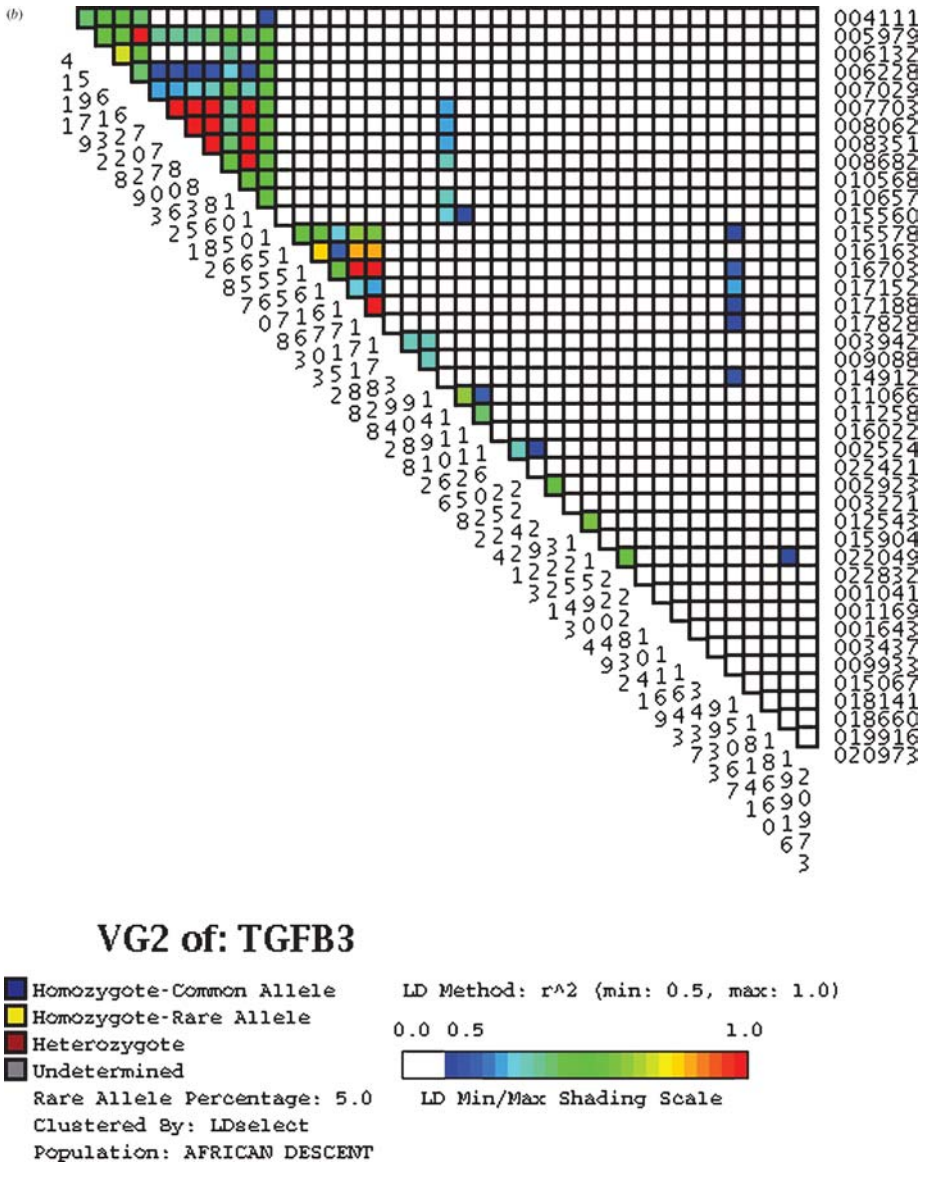

Figure 2 (*Continued*)

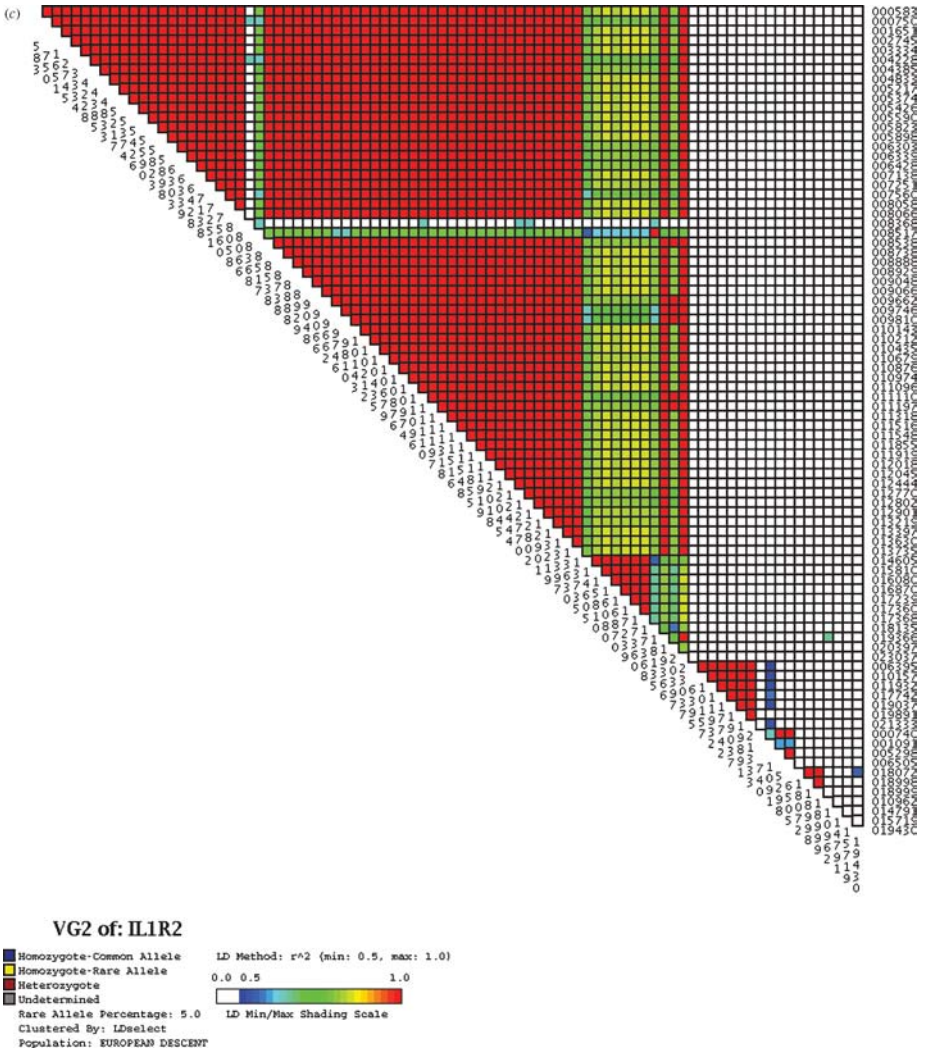

Figure 2 The strength of pair-wise linkage disequilibrium (LD) varies dramatically across candidate genes and across different population samples. Pair-wise LD is measured using VG2, software that measures LD with r^2 or D'. Here, we measured LD using r^2 for all single nucleotide polymorphisms (SNPs) with a minor allele frequency of $\geq$5%. The strength of LD is depicted graphically for each pair-wise comparison (*squares*), such that white and blue represent low levels of LD and orange and red represent high levels of LD (see color key). The numbers along the triangle represent the SNPs, which are numbered according to a reference sequence. (*a*) LD across TGFB3 in European Americans (n = 23). (*b*) LD across TGFB3 in African Americans (n = 24). (*c*) LD across IL1R2 in European Americans (n = 23).

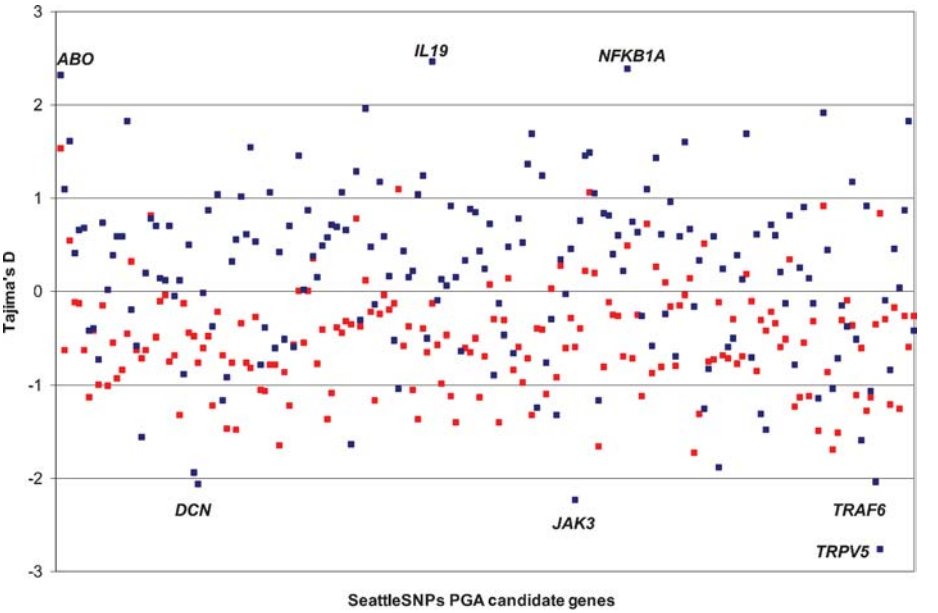

Figure 3 Tajima's D calculated across 180 SeattleSNPs candidate genes for African Americans and European Americans. Each square represents Tajima's D for a gene in the African-American sample (*red*) and the European-American sample (*blue*). Tajima's D was calculated using the methods of Tajima (103), and genes that have a Tajima's D of >2 or ≤ 2 are labeled with the gene name.

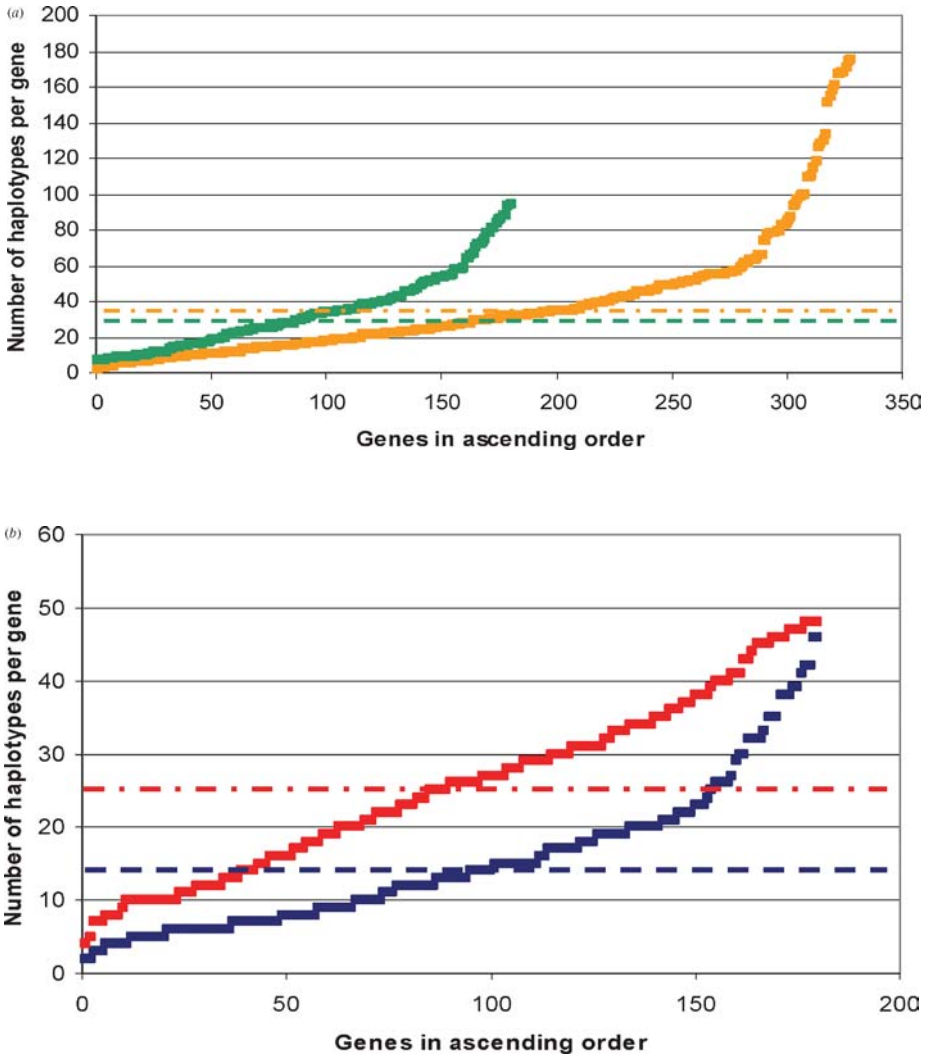

Figure 4 Haplotype diversity across 510 candidate genes. Haplotypes were inferred using PHASEv2.1 for all single nucleotide polymorphisms (SNPs) with a minor allele frequency ≥5%. (*a*) Haplotypes were inferred in the SeattleSNPs PGA samples (African and European Americans; n = 47) and EGP samples (Polymorphism Discovery Resource sample; n = 90). The dashed line represents the average number of haplotypes per gene for the SeattleSNPs data set (*green*; 34) and the EGP data set (*orange*; 38). (*b*) Haplotypes were inferred separately for African-American samples (*red*; n = 24) and European-American samples (*blue*; n = 23) in the SeattleSNPs PGA data set. The dashed line represents the average number of haplotypes per gene for African Americans (25) and European Americans (15).

Annual Review of Genomics and Human Genetics
Volume 6, 2005

CONTENTS