Hum Genet (2006) 119: 496–504
DOI 10.1007/s00439-006-0160-y

ORIGINAL INVESTIGATION

**Dana C. Crawford · Qian Yi · Joshua D. Smith
Cynthia Shephard · Michelle Wong · Laura Witrak
Robert J. Livingston · Mark J. Rieder
Deborah A. Nickerson**

Allelic spectrum of the natural variation in *CRP*

Received: 10 November 2005 / Accepted: 29 January 2006 / Published online: 21 March 2006

Abstract With the recent completion of the International HapMap Project, many tools are in hand for genetic association studies seeking to test the common variant/common disease hypothesis. In contrast, very few tools and resources are in place for genotype–phenotype studies hypothesizing that rare variation has a large impact on the phenotype of interest. To create these tools for rare variant/common disease studies, much interest is being generated towards investing in re-sequencing either large sample sizes of random chromosomes or smaller sample sizes of patients with extreme phenotypes. As a case study for rare variant discovery in random chromosomes, we have re-sequenced ~1,000 chromosomes representing diverse populations for the gene C-reactive protein (*CRP*). *CRP* is an important gene in the fields of cardiovascular and inflammation genetics, and its size (~2 kb) makes it particularly amenable medical or deep re-sequencing. With these data, we explore several issues related to the present-day candidate gene association study including the benefits of complete SNP discovery, the effects of tagSNP selection across diverse populations, and completeness of dbSNP for *CRP*. Also, we show that while deep re-sequencing uncovers potentially medically relevant coding SNPs, these SNPs are fleetingly rare when genotyped in a population-based survey of 7,000 Americans (NHANES III). Collectively, these data suggest that several different types re-sequencing and genotyping approaches may be required to fully understand the complete spectrum of alleles that impact human phenotypes.

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00439-006-0160-y and is accessible for authorized users.

D. C. Crawford (✉) · Q. Yi · J. D. Smith · C. Shephard
M. Wong · L. Witrak · R. J. Livingston · M. J. Rieder
D. A. Nickerson
Department of Genome Sciences, University of Washington,
1705 NE Pacific, Seattle, WA 98195-7730, USA
E-mail: dcrawfo@gs.washington.edu
Tel.: +1-206-6857342
Fax: +1-206-2216498

Introduction

Since the completion of the Human Genome Project, much emphasis has been placed on cataloguing human DNA variation and understanding its impact on susceptibility to human diseases. The most common form of sequence polymorphism is the single base substitution. Single nucleotide polymorphisms (SNPs) occur on average once every 180 basepairs in the human genome (Crawford et al. 2005). A catalogue of common DNA variation is now well-developed in at least one population (Carlson et al. 2003; Reich et al. 2003), and new resources provided by the International HapMap Project (The International HapMap Consortium 2003, 2005) and Perlegen (Hinds et al. 2005), among others, offer useful information and tools for genetic association studies for applying common SNPs in different population samples.

While public databases are useful for mining information about common SNPs, the databases are less useful and complete for rare variants. Rare sequence variants, defined here as having a minor allele frequency of $< 1\%$, are not well-described most likely because of the number of chromosomes required to discover them (Kruglyak and Nickerson 2001). It is possible that both common and rare variation will be required to aid investigators in developing approaches for successful genetic association study designs. For example, one common design for association studies, the common disease/common variant approach, assumes that common DNA variation rather than rare DNA variation is associated with susceptibility to common human diseases (Collins et al. 1997). However, many advocate the possibility that many rare variants, each with a small contribution, underlie the susceptibility to common human diseases (Pritchard and Cox 2002). The arguments are further complicated by study designs that genotype "tagSNPs" and rely on linkage disequilibrium to identify the causative SNP ("indirect") versus genotyping known nonsynonymous SNPs ("direct"; Botstein and Risch 2003).

As views of common variation emerge, empirical data on how to obtain rare variation and its application in the context of a genetic association study are needed in this age where "medical" re-sequencing is becoming possible (Gibbs 2005). To provide these data, we have chosen to examine these issues using the gene C-reactive protein (*CRP*). *CRP* is an ideal model for deep or medical re-sequencing because it is a small gene (~2 kb) whose protein product is well studied as an established general marker for predicting cardiovascular disease and risk of myocardial infarction (Danesh et al. 2004). Furthermore, sequence variation for *CRP* has been described in several smaller studies (Brull et al. 2003; Cao and Hegele 2000; Kovacs et al. 2005; Russell et al. 2004; Wolford et al. 2003), only one of which has identified nonsynonymous variation within the gene (Miller et al. 2005). Here, we describe the re-sequencing of a single gene, *CRP*, for the purpose of cataloguing both common and rare variation, and we also describe the possible application of the discovered variants in population-based survey using both "indirect" and "direct" association study designs. We demonstrate that while deep re-sequencing or re-sequencing in several racial/ethnic populations results in the identification of rare, potentially damaging nonsynonymous SNPs, the effort may not necessarily translate to a better understanding of the gene's impact on human phenotypes when applied to larger cohorts linked to quantitative phenotypes.

Materials and methods

Sequencing

As part of NHLBI's Program for Genomic Applications, SeattleSNPs re-sequences genes involved in inflammation, lipid metabolism, and blood pressure regulation for genetic variation discovery in two population samples: European-Americans ($n = 23$) and African-Americans ($n = 24$) obtained from Coriell Cell Repository. The European-American samples are members of the CEPH panel (NA12560, NA12547, NA10845, NA10853, NA10860, NA10830, NA10842, NA10851, NA07349, NA10857, NA10858, NA10848, NA12548, NA10844, NA10854, NA10861, NA10831, NA10843, NA10850, NA07348, NA10852, NA06990, NA07019) and the African-American samples are members of the African-American panel of 50 (NA17101-NA17116; NA17133-NA17140). Additional sequencing for variation discovery was performed on samples from the Polymorphism Discovery Resource Panel (PDR450)(Collins et al. 1998) and several other population samples obtained from Coriell Cell Repository, including seven Chinese (NA16654, NA16688-NA16689, NA17014-NA17017), seven Japanese (NA17051-NA17057), ten Southeast Asians (NA17081-NA17090), ten Mexicans (NA17061-NA17070), and six Indo-Pakistanis (NA17021-NA17024; NA17026-NA17027).

All samples were re-sequenced for variation discovery in *CRP* on an ABI3730 using standard Big Dye terminator chemistry. A detailed laboratory protocol can be found on the SeattleSNPs website (see Websites). A total of 6.8 kb was re-sequenced for *CRP*, including all exons, all introns, 1.7 kb 5′ flanking, and 2.8 kb 3′ flanking. SNPs were numbered based on the GenBank accession number AF449713 and were submitted to dbSNP. The SNP number does not correspond with the translation start site. The location, rs number (where available), and sequence context of the SNPs described here can be found in Supplementary Table 1. All SNP allele frequencies and 95% confidence intervals for each population sample are given in Supplementary Table 2.

Haplotypes were inferred using PHASEv2.1.1 (Stephens et al. 2001; Stephens and Donnelly 2003) using SNPs with a minor allele frequency of $> 5\%$. Inferred haplotypes and their counts are given in Supplementary Table 3 for each population sample. TagSNPs were determined using LDSelect with a minor allele frequency $> 5\%$ and $r^2 > 0.64$ (Carlson et al. 2004). The predicted effects of all nonsynonymous variation were explored computationally using PolyPhen (Ramensky et al. 2002) and SIFT (Ng and Henikoff 2002).

Genotyping

Standard TaqMan® Assays by Design® were developed to genotype two sites newly discovered in *CRP*: 2513 and 2612. The sequence for the primers and Taqman® probes for 2513 was 5′-GATCGTGGAGTTCTGGG TAGATG-3′ (forward), 5′-ACAGTGTATCCCTTCT TCAGACTCT-3′ (reverse), 5′-CACCCTGGGCTTC-3′ (C allele probe), and 5′-TCACCCTGAGCTTC-3′ (T allele probe). For 2612, the primers and probes were 5′-GGGCAGGAGCAGGATTCC-3′ (forward), 5′-AG TCCCACATGTTCACATTTCCAAT-3′ (reverse), 5′-AC TTTGAAGGAAGCCAG-3′ (G allele probe), and 5′-ACTTTGAAGAAAGCCAG-3′ (A allele probe). The two assays were validated, and individuals identified as heterozygous for 2612 were confirmed by re-sequencing.

DNA samples ($n = 7{,}296$) from the National Health and Nutrition Examination Survey (NHANES) III were genotyped for the two sites by Taqman®. The genotyping success rate was 97% for both SNPs. NHANES III is a population-based sample of Americans collected regardless of health status between 1988 and 1994, and DNA was collected for phase 2 of the survey (1991–1994). All DNA samples are linked to an extensive collection of laboratory phenotypes, such as serum CRP levels, as well as survey data. Survey design and ascertainment of NHANES III as well as descriptions of the phenotypic variables can be found at CDC's National Center for Health Statistics website (see Websites). The three largest racial/ethnic groups for NHANES III include non-Hispanic whites ($n = 2{,}630$), non-Hispanic blacks ($n = 2{,}108$), and Mexican-Americans ($n = 2{,}073$). A small number of samples ($n = 137$) were

duplicates or not linked to the phenotypic data and were not included in subsequent analyses.

Results

SNP discovery

Three sets of population samples were re-sequenced for variation discovery in *CRP* with different goals: (1) the standard SeattleSNPs population samples re-sequenced for mostly common variation discovery in European-Americans ($n = 23$) and African-Americans ($n = 24$); (2) an extended DNA panel of Asians (Chinese, $n = 7$; Japanese, $n = 7$, Southeast Asian, $n = 10$), Mexicans ($n = 10$), and Indo-Pakistanis ($n = 6$) for population-specific variation discovery; and (3) the Polymorphism Discovery Resource Panel (PDR; $n = 450$) for rare coding variation discovery (Collins et al. 1998). Both the SeattleSNPs and extended DNA panels have sufficient power to identify common SNPs (minor allele frequency $> 5\%$); however, the range of power depends on the number of chromosomes re-sequenced for each population sample. Thus, the African-American and combined Asian samples have the highest detection rate for common SNPs (99%), and the Indo-Pakistani sample has the lowest detect rate for common SNPs ($> 80\%$)(Kruglyak and Nickerson 2001). The PDR can detect 99% of SNPs with a MAF of at least 1% (Kruglyak and Nickerson 2001).

To determine whether SeattleSNPs is capturing all *CRP* common variation found in many populations by re-sequencing only European- and African-American samples, we compared the SNPs identified by the usual SeattleSNPs DNA panel to the extended DNA panel containing samples of Asian, Mexican, and Indo-Pakistani ancestry. From these two DNA panels, a total of 40 SNPs were identified (Fig. 1 and Table 1). As expected, a greater number of SNPs was identified in the African-American population sample (30) compared with the Asian (17), Mexican (13), European-American (13), and Indo-Pakistani (11) population samples (Table 1). Also, an increase in nucleotide diversity as well as a greater number of haplotypes was observed in the African-American population samples compared with the other population samples (Table 1). Approximately half (58%) of the 40 SNPs are private to one population sample, most of which were found in the African-American population sample (14 SNPs). The inclusion of the extended DNA panel allowed the identification of nine SNPs not identified in the SeattleSNPs panel. Two of the nine new sites were common ($> 5\%$ MAF) within the Indo-Pakistani population sample (sites 974 and 2513).

Among the 40 SNPs identified in *CRP*, one unique variant is the high frequency triallelic SNP 1440 (rs3091244). All three alleles of 1440, A, C, and T, occur with appreciable frequency in the African-American (0.33, 0.26, and 0.41) and European-American (0.05, 0.62, and 0.33) population samples. Examination of the extended DNA panel reveals that the combined Asian population sample of Chinese, Japanese, and Southeast Asian samples is also polymorphic for all three 1440 alleles (0.19, 0.75, and 0.06); however, the allele frequency distribution in the combined Asian population sample is very different compared with the African- and European-American population samples. The Mexican population sample does not contain all three alleles, possibly due to the smaller sample size of this collection compared with the other three population samples. Among chromosomes of Mexican-descent, the C allele occurs more frequently (0.75) than the T allele. Finally, among the Indo-Pakistani samples, the C allele is predominant over the A allele (0.83 vs. 0.17). Again, the absence of the third allele (T) may be due to the small sample size of this population collection compared with the larger collections.

Coding variation

One of the sites identified in the Indo-Pakistani sample, SNP 2513 (MAF = 0.08), is a nonsynonymous SNP (Pro133Lys) in exon 2 of *CRP*. Site 2513 is predicted to be "possibly damaging" by PolyPhen but "tolerated" by SIFT (Table 2a). Also, site 2475 identified in two heterozygous individuals from the Japanese population sample is a novel synonymous SNP in exon 2 (Table 2a).

Given the possibility that *CRP* contains previously unidentified rare and potentially population-specific nonsynonymous SNPs, we re-sequenced *CRP* in individuals from the Polymorphism Discovery Resource Panel (PDR450) primarily to discover novel coding variation. The PDR450 samples are not linked to racial/ethnic identifiers; therefore, the panel is useful for DNA variation discovery, but not for allele frequency estimation for specific populations. Overall, 46 sites were identified in the PDR, 61% of which were not found

Table 1 Sample size (chromosomes), number of segregating sites (S), nucleotide diversity (θ), and number of inferred haplotypes for *CRP* re-sequenced in several samples

Sample	N	S (MAF > 5%)	θ (10^{-4})	Haplotypes (heterozygosity)
African-Americans	48	30 (18)	11.96	18 (0.86)
European-Americans	46	13 (10)	5.25	7 (0.76)
Asians	48	17 (9)	6.84	11 (0.67)
Mexicans	20	13 (11)	6.32	6 (0.74)
Indo-Pakistanis	12	11 (11)	7.17	6 (0.75)
Total	174	40 (17)	11.83	29 (0.86)

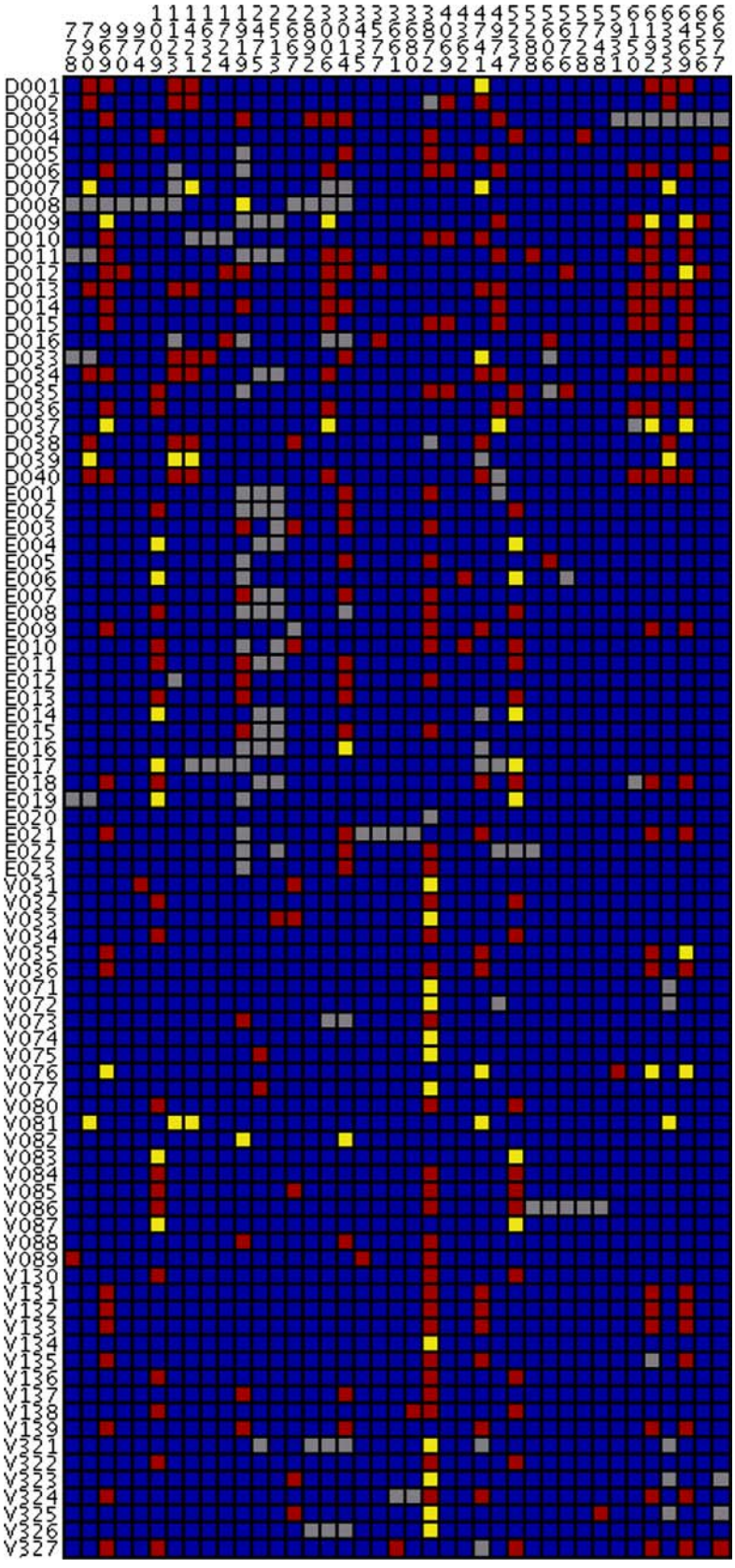

Fig. 1 *CRP* variants identified re-sequencing the SeattleSNPs ($n=47$) and extended DNA panels ($n=40$). SNPs are numbered across the top of the figure in order of discovery in the reference sequence. Individual samples are labeled to the left of the figure. Each *square* represents the genotype of a SNP for each individual: homozygous common (*blue*), heterozygous (*red*), and homozygous rare (*yellow*). *Gray* represents missing data. *Abbreviations*: African-American (*D*), European-American (*E*), extended panel DNA (*V31-36*: Indo-Pakistani, *V80-89*: Mexican; and *V71-77, V130-139, V321-327*: Asian)

among the SeattleSNPs and extended DNA panel samples. All novel sites in the PDR had a minor allele frequency of 2% or less (Supplementary Table 2). Among these novel sites, we identified two additional nonsynonymous SNPs, sites 2314 (Tyr67His) and 2612 (Gly166Glu), each in a single heterozygous individual. Site 2314 is predicted

to be tolerated/benign and site 2612 is predicted to be tolerated/possibly damaging by SIFT and PolyPhen, respectively (Table 2a). A previously described (site 2667) as well as two novel synonymous SNPs (sites 2220 and 2244) was also identified in the PDR (Table 2b). Not surprisingly, the synonymous sites, unlike the nonsynonymous sites, were identified in more than one individual. The nonsynonymous site 2513 and the synonymous site 2475 identified in the Indo-Pakistani and Japanese samples, respectively, were not identified in the larger PDR sample. We could not confirm the recently reported nonsynonymous Thr45Met in any of the populations re-sequenced here (Miller et al. 2005).

The newly identified nonsynonymous SNPs are of obvious interest to investigators designing direct genetic association studies because it is assumed that these SNPs are more likely to have a direct impact on the phenotype than noncoding SNPs (Botstein and Risch 2003). However, nonsynonymous SNPs are only informative if they occur at an appreciable frequency in the general population or are enriched among individuals with specific diseases/outcomes. To better characterize the allele and genotype frequency of the possibly damaging nonsynonymous SNPs in a general population, we genotyped sites 2513 and 2612 in a sample of 7,296 DNAs collected by the Centers for Disease Control and Prevention for phase 2 of the National Health and Nutrition Examination Survey (NHANES) III. In $>14{,}000$ chromosomes examined, no individual was heterozygous for site 2513. For site 2612, we identified two heterozygous individuals. Thus, these two sites are rare in the general population with a minor allele frequency well below 1%.

CRP and dbSNP

For both direct and indirect association study designs, one common concern for investigators is the reliance on public databases for available genetic variations in the gene or region of interest. Previous work examining the quality and completeness of dbSNP suggests that $>50\%$ of the reported SNPs are common (Carlson et al. 2003; Jiang et al. 2003; Reich et al. 2003). These studies also demonstrate that a considerable fraction of SNPs in dbSNP ($\sim$40%) cannot be found in other samples, suggesting these reported SNPs are either very rare or artifacts (Carlson et al. 2003; Jiang et al. 2003). To examine these issues for *CRP*, we mapped the SNPs identified here by deep re-sequencing to those found within dbSNP in the gene *CRP*.

For build 125 of dbSNP, 17 DNA variations are reported in the gene *CRP*. Five of the 17 DNA variations reported by dbSNP are considered validated SNPs. To directly compare SNPs reported in dbSNP to the SNPs discovered in the various populations reported here, we only considered the 21 SNPs we annotated in exons 1 and 2 and intron 1. Only six of the 17 dbSNP reports (35%) can be confirmed by the re-sequencing efforts described here, and all six of these SNPs were common SNPs reported by SeattleSNPs. These six SNPs included the five validated SNPs. More than half of the dbSNP submissions (11/17; 64.7%) cannot be confirmed by any of the populations re-sequenced here. Overall, 71.4% of the *CRP* variants identified here are not currently found in dbSNP; however, none of these SNPs is common among the chromosomes re-sequenced here. Thus, dbSNP is most likely complete for common variation but incomplete for rare variation in *CRP*, as expected.

TagSNPs for association studies

Another common concern for investigators is that an optimal subset of SNPs (termed "tagSNPs") must be chosen from the information at hand for genotyping in a genetic association study. The tagSNP or "indirect" approach can be an economical approach because only SNPs that maintain the genetic diversity of the gene or region (as defined by linkage disequilibrium) are genotyped (Carlson et al. 2004; Stram 2004). Recent evidence suggests that tagSNPs are transferable across populations of similar race/ethnicity without loss of informa-

Table 2 *CRP* variants identified in the coding region

Site	Nucleotide change	Amino acid change (position)	Number of chromosomes (panel)	SIFT/PolyPhen prediction
(a) Nonsynonymous				
2314	T > C	Tyr to His (67)	1/794 (PDR)	Tolerated/benign
2513	C > T	Pro to Lys (133)	1/80 (extended)	Tolerated/possibly damaging
2612	G > A	Gly to Glu (166)	1/828 (PDR)	Tolerated/possibly damaging
(b) Synonymous				
2220	T > G	Thr to Thr (35)	7/700 (PDR)	N/A
2244	G > A	Pro to Pro (43)	3/734 (PDR)	N/A
2475	C > T	Ser to Ser (120)	2/80 (extended)	N/A
2667 (rs1800947)	G > C	Leu to Leu (184)	36/826 (PDR)	N/A

PDR Polymorphism Discovery Resource panel

tion (Ke et al. 2004). SeattleSNPs routinely chooses tagSNPs for European-American and African-American population samples, and it is expected that these tagSNPs will be sufficient for association studies genotyping samples of European- and African-descent. However, it is not know how well these tagSNPs perform when genotyped in an admixed population such as Mexican-Americans.

Previous admixture studies suggest that Mexican or Mexican-American populations have European, African, and Asian/Amerindian ancestry, with proportions estimated at approximately 40–60, 3–5, and 30–50%, respectively (Hanis et al. 1991; Long et al. 1991; Collins-Schramm et al. 2004; Salari et al. 2005). To understand how much, if any, diversity information is lost in genotyping SeattleSNPs tagSNPs in a Mexican or Mexican-American sample, we compared tagSNPs determined for a combined population of African- and European-Americans (Fig. 2a, b) to those determined for the Mexican population sample (Fig. 2c) using LDSelect at MAF > 5% and $r^2 > 0.64$. For the Mexican population sample of 20 chromosomes, five tagSNPs were identified for genotyping. All five tagSNPs were represented among tagSNPs chosen for the

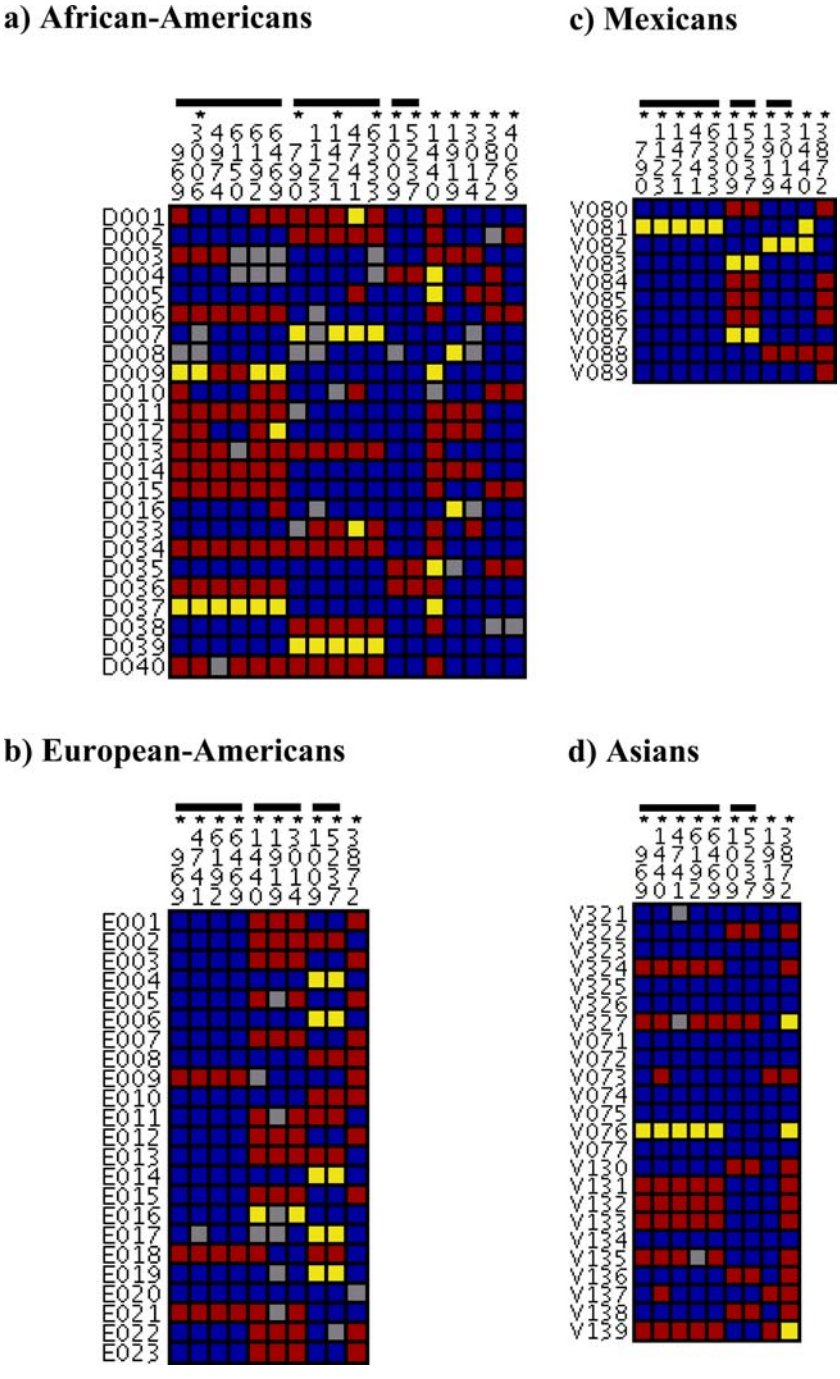

Fig. 2 TagSNPs chosen for four population samples. TagSNPs were chosen from SNPs with MAF > 5% using LDSelect at $r^2 > 0.64$

European- and African-American combined sample and are sufficient to capture the genetic diversity of the locus within the Mexican population sample. Interestingly, the Asian population sample of Chinese, Japanese, and Southeastern Asians is also well represented by *CRP* tagSNPs chosen from a combined European- and African-American population sample (Fig. 2d).

Conclusions

By re-sequencing ~1,000 chromosomes ascertained from several population samples, we have catalogued novel coding and noncoding DNA variants in *CRP*. The vast majority of these novel SNPs were rare, consistent with previous observations from candidate gene deep re-sequencing projects (Glatt et al. 2001). We also identified common *CRP* variation in the Asian and Mexican-descent population samples that overlapped to a great extent with common variation identified by SeattleSNPs in the European-American and African-American population samples.

For *CRP*-specific genetic association studies, these re-sequencing data suggest that *CRP* common SNPs are in dbSNP and that common disease/common variant study designs using a tagSNP approach can be applied to several population samples with very little population-specific modification. That is, if the most diverse population (e.g., African-descent populations) is considered in tagSNP selection for *CRP*, genetic diversity for *CRP* in other population samples will be well represented. However, if only a less diverse population were considered for tagSNP selection (e.g., European-descent populations), one tagSNP (rs3093058) associated with serum CRP levels (Carlson et al. 2005; D. C. Crawford et al., submitted) would not be represented for *CRP* diversity observed in African-Americans and Mexican-Americans. This issue should be considered when designing genetic association studies in non-European population samples using public SNP resources such as HapMap or dbSNP (Clark et al. 2005).

The re-sequencing data for *CRP* also suggest that coding variation, if not identified in a small re-sequencing sample, is very rare in the general population. These results are not entirely unexpected because the samples sizes used in variation discovery for the present study are based on population genetics theory (i.e., effective population size and mutation rate) and are designed to sufficiently estimate the allele frequency for SNPs with a minor allele frequency of >5–10% (Kruglyak and Nickerson 2001). Indeed, tagSNPs determined in SeattleSNPs were recently genotyped in three larger studies and demonstrated that allele frequencies estimated in the SeattleSNPs sample of 47 individuals is nearly identical to the allele frequencies estimated in the larger studies (Carlson et al. 2005; D. C. Crawford et al., submitted; L. A. Lange et al., in preparation).

While the coding variation identified in the present study was rare in the general population, there is a possibility that *CRP* coding variation is strongly associated with a phenotype and could be enriched in a case population. The challenge, of course, is to predict a disease or extreme phenotype given that a putative mutation has been identified. Given that the nonsynonymous SNP in *CRP* may be damaging to the gene product, we may expect lower levels of serum CRP in heterozygous individuals compared with homozygous individuals for the common allele. Alternately, we may expect individuals heterozygous for these SNPs are unable to mount a proper acute phase response compared with individuals homozygous for the common allele. There is no known human deficiency for CRP, but there are several phenotypes associated with lack of acute phase response (reviewed in Pepys and Hirschfield 2003). Unfortunately, the CRP assay used in NHANES III is not as sensitive compared with currently used assays (Centers for Disease Control and Prevention 1994), so we cannot determine how low the serum CRP levels are for the two individuals who are heterozygous for the nonsynonymous *CRP* SNP (D. C. Crawford et al., submitted).

The possibility that deleterious variation may exist in human *CRP* is intriguing. In addition to providing information for future functional *CRP* studies in humans, these deep re-sequencing data provide an opportunity to explore the effects of possibly deleterious genetic variation using model organisms and in vitro assays. One obvious choice for modeling SNP effects in context with various environmental exposures is the mouse (MacAuley and Ladiges 2005). However, for *CRP*, the mouse is not an ideal model because its *CRP* gene does not function like the human *CRP* gene (Pepys and Hirschfield 2003). Nevertheless, the *CRP* example demonstrates the potential usefulness of identifying coding variants through deep re-sequencing that may be taken forward into other meaningful experiments that specialize in function.

In summary, we describe here the genetic architecture of *CRP* with special emphasis on the detection and characterization of rare variants. While we were able to describe novel nonsynonymous variants for *CRP*, we were unable to describe these variants at an appreciable frequency when genotyped in the general population making it impossible to explore the impact of these variants on serum CRP. Given this situation, for a quantitative trait such as CRP, it may be more efficient to re-sequence individuals at the extreme low and high end of the distribution in the search for rarer SNPs that impact the phenotype of interest (Cohen et al. 2005; Cohen et al. 2004). In the age of more efficient, high throughput genotyping (Hinds et al. 2004; Shen et al. 2005), it may be that re-sequencing used in conjunction with genotyping using different study designs is required to fully appreciate the spectrum of genetic variants associated with a human phenotype of interest.

Websites

EntrezGene http://www.ncbi.nih.gov/entrez/query.fcgi?db=gene

NHANES: http://www.cdc.gov/nchs/nhanes.htm

PolyPhen: http://www.tux.embl-heidelberg.de/ramensky/

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

SIFT: http://www.blocks.fhcrc.org/sift/SIFT_BLink_submit.html

Acknowledgements We thank C.L. Sanders and Dr. X. Qin (NCHS/CDC; Harris Corporation at Falls Church, Virginia) for their assistance on calculating the allele frequency for SNP 2616 genotyped in NHANES III. The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the Centers for Disease Control and Prevention. This work was supported by grants N01 ES15478 and U01 HL66682.

References

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33(Suppl):228–237

Brull DJ, Serrano N, Zito F, Jones L, Montgomery HE, Rumley A, Sharma P, Lowe GDO, World MJ, Humphries SE, Hingorani AD (2003) Human CRP gene polymorphism influences CRP levels: implications for the prediction and pathogenesis of coronary heart disease. Arterioscler Thromb Vasc Biol 23:2063–2069

Cao H, Hegele RA (2000) Human C-reactive protein (CRP) 1059G/C polymorphism. J Hum Genet 45:100–101

Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat Genet 33:518–521

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106–120

Carlson CS, Aldred SF, Lee PK, Tracy RP, Schwartz SM, Rieder M, Liu K, Williams OD, Iribarren C, Lewis EC, Fornage M, Boerwinkle E, Gross M, Jaquish C, Nickerson DA, Myers RM, Siscovick DS, Reiner AP (2005) Polymorphisms within the C-reactive protein (CRP) promoter region are associated with plasma CRP levels. Am J Hum Genet 77:64–77

Centers for Disease Control and Prevention (1994) Plan and operation of the third National Health and Nutrition Examination Survey, 1988–1994. Bethesda, MD

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15:1496–1502

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305:869–872

Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH (2005) Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet 37:161–165

Collins FS, Guyer MS, Charkravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278:1580–1581

Collins FS, Brooks LD, Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 8:1229–1231

Collins-Schramm HE, Chima B, Morii T, Wah K, Figueroa Y, Criswell LA, Hanson RL, Knowler WC, Silva G, Belmont JW, Seldin MF (2004) Mexican American ancestry-informative markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. Hum Genet 114(3):263–271

Crawford DC, Akey DT, Nickerson DA (2005) The patterns of natural variation in human genes. Annu Rev Genomics Hum Genet 6:287–312

Danesh J, Wheeler JG, Hirschfield GM, Eda S, Eiriksdottir G, Rumley A, Lowe GDO, Pepys MB, Gudnason V (2004) C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease. N Engl J Med 350:1387–1397

Gibbs R (2005) Deeper into the genome. Nature 437:1233–1234

Glatt CE, DeYoung JA, Delgado S, Service SK, Giacomini KM, Edwards RH, Risch N, Freimer NB (2001) Screening a large reference sample to identify very low frequency sequence variants: comparisons between two genes. Nat Genet 27:435–438

Hanis CL, Hewett-Emmett D, Bertin TK, Schull WJ (1991) Origins of U.S. Hispanics. Implications for diabetes. Diabetes Care 14(7):618–627

Hinds DA, Stokowski RP, Patil N, Konvicka K, Kershenobich D, Cox DR, Ballinger DG (2004) Matching strategies for genetic association studies in structured populations. Am J Hum Genet 74:317–325

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079

Jiang R, Duan J, Windemuth A, Stephens JC, Judson R, Xu C (2003) Genome-wide evaluation of the public SNP databases. Pharmacogenomics 4:779–789

Ke X, Durrant C, Morris AP, Hunt S, Bentley DR, Deloukas P, Cardon LR (2004) Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. Hum Mol Genet 13(21):2557–2565

Kovacs A, Green F, Hannson L-O, Lundman P, Samnegard A, Boquist S, Ericsson C-G, Watkins H, Hamsten A, Tornvall P (2005) A novel common single nucleotide polymorphism in the promoter region of the C-reactive protein gene associated with the plasma concentration of C-reactive protein. Atherosclerosis 178:193–198

Kruglyak L, Nickerson DA (2001) Variation is the spice of life. Nat Genet 27:234–236

Long JC, Williams RC, McAuley JE, Medis R, Partel R, Tregellas WM, South SF, Rea AE, McCormick SB, Iwaniec U (1991) Genetic variation in Arizona Mexican Americans: estimation and interpretation of admixture proportions. Am J Phys Anthropol 84(2):141–157

MacAuley A, Ladiges WC (2005) Approaches to determine clinical significance of genetic variants. Mutat Res 573:205–220

Miller DT, Zee RYL, Danik JS, Kozlowski P, Chasman DI, Lazarus R, Cook NR, Ridker PM, Kwiatkowski DJ (2005) Association of common CRP gene variants with CRP levels and cardiovascular events. Ann Hum Genet 69:623–638

Ng PC, Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. Genome Res 12:436–446

Pepys MB, Hirschfield GM (2003) C-reactive protein: a critical update. J Clin Invest 111:1805–1812

Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet 11:2417–2423

Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894–3900

Reich DE, Gabriel SB, Altshuler D (2003) Quality and completeness of SNP databases. Nat Genet 33:457–458

Russell AI, Cunninghame Graham DS, Shepherd C, Roberton CA, Whittaker J, Meeks J, Powell RJ, Isenberg DA, Walport MJ, Vyse TJ (2004) Polymorphism at the C-reactive protein locus influences gene expression and predisposes to systemic lupus erythematosus. Hum Mol Genet 13:134–147

Salari K, Choudhry S, Tang H, Naqvi M, Lind D, Avila PC, Coyle NE, Ung N, Nazario S, Casal J, Torres-Palacios A, Clark S,

Phong A, Gomez I, Matallana H, Pérez-Stable EJ, Shriver MD, Kwok P-Y, Sheppard D, Rodriguez-Cintron W, Risch NJ, Burchard EG, Ziv E (2005) Genetic admixture and asthma-related phenotypes in Mexican American and Puerto Rican asthmatics. Genet Epidemiol 29(1):76–86

Shen R, Fan J-B, Campbell D, Chang W, Chen J, Doucet D, Yeakley J, Bibikova M, Garcia EW, McBride C, Steemers F, Garcia F, Kermani BG, Gunderson K, Oliphant A (2005) High-throughput SNP genotyping on universal bead arrays. Mutat Res 573:70–82

Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169

Stephens M, Smith NJ, Donnelly P (2001) A new statstical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Stram DO (2004) Tag SNP selection for association studies. Genet Epidemiol 27:365–374

The International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

Wolford JK, Gruber JD, Ossowski VM, Vozarova B, Antonio TP, Bogardus C, Hanson RL (2003) A C-reactive protein promoter polymorphism is associated with type 2 diabetes mellitus in Pima Indians. Mol Genet Metab 78:136–144