An Overview of the Environmental Genome Project

Deborah A. Nickerson, Mark J. Rieder, Dana C. Crawford, Christopher S. Carlson, and Robert J. Livingston

From left to right: Christopher Carlson, Deborah Nickerson, Mark Rieder, Diana Crawford, Robert Livingston.

All co-authors are affiliated with the Department of Genome Sciences, University of Washington, Seattle, Washington.

Providing a Resource to Explore Phenotype–Genotype–Environment Interactions

Understanding the causes of common diseases such as cancer, asthma, diabetes, hypertension, and atherosclerosis, which have high population prevalence, is a significant priority for public health research and a major goal in biomedical studies. The influences on susceptibility to common disease are thought to arise from multiple factors, each conferring a low level of relative risk for disease. These low levels of disease risk probably reflect interactions between genotypes at multiple loci (epistasis), interactions between genotypes and environment, and more stochastic epigenetic events such as methylation (Belinsky 2004; Berwick 2000; Hayward 2003; Hegele 1997). Because the risk of any given effect is small (Lander and Schork 1994; Moffatt and Cookson 1999; Pritchard and Cox 2002; Reich and Lander 2001; Risch and Merikangas 1996), detecting these influences will require larger sample sizes, making population-based association studies more practical for tracing the underlying genetic risks (Risch and Merikangas 1996). Association studies will also be key for exploring genetic links to environmental exposures (Bell and Taylor 1997).

In 1997 Dr. Kenneth Olden, director of the National Institute of Environmental Health Sciences (NIEHS), convened a historic conference titled "The Environmental Genome Project" held 17–18 October 1997 in Bethesda, Maryland. This symposium explored the feasibility of the Environmental Genome Project (EGP) and generated significant discussion. The EGP is designed to explore the relationship between common genetic polymorphisms and environmentally induced disease in human populations

From its inception, the Environmental Genome Project has provided the genotype data needed to drive the next-generation association mapping of genotype-phenotype-environmental interactions. This project has revealed the broad gene-to-gene variation in sequence diversity, linkage disequilibrium, and haplotype diversity present in the human genome.

(Olden and Wilson 2000). This key symposium focused on themes that even today remain of great importance: *a*) the known interactions between genetic variants, environmental agents, and disease risk; *b*) the current and emerging technologies to identify and type DNA polymorphisms in the human genome; *c*) sequence diversity and human population genetics; and *d*) the available functional tools to analyze DNA polymorphisms.

The discussion and follow-up to this 1997 conference have had a profound impact on human genetic analysis and have formed the foundation of the EGP as well as many other large-scale projects aimed at defining the variability of the human genome (Collins et al. 1997; Olden and Wilson 2000) and studying gene–environment interactions (Collins et al. 2003). In this overview, we discuss progress in the EGP and prospects for the future in terms of the themes envisioned by Dr. Olden.

Human Disease and Gene–Environment Interactions

More than a century ago the link between environmental exposure and disease susceptibility was first recognized with the discovery of the association between exposure to coal soot and cancer in young chimney sweeps (Doll 1975). Other early examples developed from typing protein polymorphisms in human populations. These include hemolysis in individuals with glucose 6-phosphate dehydrogenase deficiency after exposure to antimalarial drugs and other oxidants (Motulsky 1972); increased risk of emphysema from cigarette smoking in individuals with α_1 antitrypsin deficiency (Eriksson 1965; Lieberman et al. 1969) and, lactose intolerance in individuals with lactase deficiency (Dahlqvist et al. 1963; Haemmerli et al. 1965; Klotz 1964). Over the past three decades, numerous links between DNA variations in the enzymes that metabolize and/or detoxify carcinogens and susceptibility to specific cancers with exposure to environmental agents have been reported (Kelada et al. 2003). In the future new technologies for directly quantifying environmental exposures will be required to improve the accuracy of gene–environment associations (Rothman et al. 1999). New system-based approaches such as proteomic and metabolic profiling will likely play a central role in these analyses and will provide quantitative data on environment exposures. These are now being implemented into studies of toxigenomics and the EGP (Waters and Fostel 2004).

Identifying Single Nucleotide Polymorphisms and the EGP

The links between environmental exposures and polymorphism analysis have a long history, and association studies are clearly suited to approaching these analyses (Bell and Taylor 1997). Association studies compare the frequency of a polymorphic marker, or a set of markers, in affected and unaffected individuals. Because recombination along the chromosomes is averaged over the genetic history of the population at large, this should randomize any association between a given polymorphism and phenotype, or environmental influence, unless it is closely linked with specific alleles in the genome. Although association studies have greater power to detect variants with low relative risk, whole-genome association studies will require extremely dense sets of polymorphic markers—on the order of hundreds of thousands to more than a million markers that can rapidly be typed on large numbers of samples (Kruglyak 1999).

To develop such high-density genetic maps, studies have focused on the identification of single-nucleotide substitutions because they are the most abundant form of sequence variation in the human genome (Cooper et al. 1985; Kruglyak 1997; Wang et al. 1998). If one considers the size of the human population (~ 6 billion), with a mutation rate of approximately 2×10^{-8} per base pair per generation, then every site in the genome compatible with survival has mutated an average of 240 times in just the most recent generations. However, most of these base substitutions are extremely rare in human populations. Only a fraction of the variation that exists has minor allele frequencies (MAF) exceeding 1%, and these are referred to as single nucleotide polymorphisms (SNPs). Recent estimates predict that > 15 million SNPs with MAFs exceeding 1%, and > 7 million markers with MAFs exceeding 5%, will be found in the human genome (Kruglyak and Nickerson 2001).

The discovery of SNPs in the human genome has been aided by the development of a panel of samples known as the polymorphism discovery resource (PDR) panel proposed by the National Human Genome Research Institute (Collins et al. 1998). The PDR panel was designed to discover human genetic variation while being sensitive to the ethical, legal, and social issues of population definition and not to assess the frequency of variations in specific ethnic subpopulations. Therefore, all identifying demographic information was removed from the individual samples. However, samples in this panel are representative of individuals drawn from the U.S. population, including Americans of European, African, Mexican, and Asian descent and Native Americans. Until recently, variation discovery in the EGP has focused on the PDR panel of 90 samples. This sample size is sufficient to detect polymorphic sites occurring at > 5% MAF in any one of the ethnic subpopulations (Kruglyak and Nickerson 2001). This focused the discovery efforts on the identification of common polymorphisms for association studies.

A number of strategies have been used to identify SNPs, and of these, DNA sequencing has become the dominant technology. To date, > 10 million SNPs have been uniquely mapped on the human genome (build 123; http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp). Most variants in the current database have been identified as single-base mismatches by comparing sequences from overlapping BAC (bacterial artificial chromosome) clones that were sequenced for the human genome or by comparing the reference genome sequence with sequences obtained by shotgun sequencing (Altshuler et al. 2000b; Sachidanandam et al. 2001; Venter et al. 2001). Frequency information is available for only a subset of these SNPs, although this is rapidly changing with the emergence of the HapMap data set (http://www.hapmap.org/) and the Perlegen data sets (Patil et al. 2001; http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp).

A surrogate strategy that has emerged to identify common SNPs in the absence of real frequency data is to rely on SNPs identified by two independent discoveries for each of the two alleles (Gabriel et al. 2002; Reich et al. 2003). These are being referred to as "double-hit" SNPs, and several recent analyses have shown these variants from the database are likely to have MAFs sufficient to be detected again in another population survey (Carlson et al. 2003; Reich et al. 2003). However, as many studies have shown, the patterns of variation in the genome are influenced by a number of factors, and the analysis of these patterns will require genotype data for each site as well as its relationship to its surrounding sites (Carlson et al. 2003; Wang and Todd 2003). The availability of comprehensive genotype information greatly aids in the selection of the most useful SNP markers for large-scale genotyping. Since its inception, the EGP has focused on generating nearly complete genotype information using targeted DNA sequencing of genes across 90 PDR samples and has provided substantial insights in the variability of the human genome (Livingston et al. 2004).

EGP Candidate Genes

After the initial EGP symposium, NIEHS investigators provided substantial input into the development of a list of 550 candidate environmental response genes for targeted variation discovery. These candidates include genes involved in DNA repair, apoptosis, cell cycle control, and drug metabolism (for the complete list, see GeneSNPs at http://genome.utah.edu/genesnps). These candidates are distributed across all the human chromosomes except for the Y-chromosome, and altogether represent > 2% of all known human genes (International Human Genome Sequencing Consortium 2004). Efforts for the EGP have also focused on completing SNP discovery across entire pathways of interacting genes, such as the base-excision repair pathway illustrated in Figure 1. Other pathways include the nucleotide-excision repair, mismatch repair,

double-stranded break repair, and transcription-coupled repair pathways. Many members of these pathways have been implicated in cancer susceptibility (Han et al. 2004; Ide and Kotera 2004; Mohrenweiser et al. 2002).

SNP Discovery in the EGP

To date, the discovery efforts for the EGP represent the largest resequencing effort ever attempted across the human genome. A total of 371 genes have been scanned, and on average, approximately 53% of the genomic sequence for each gene has been examined for variation across the 90 PDR samples. Notably, approximately 20% of these candidate genes have already been implicated in Mendelian diseases, including disease genes for rare forms of cancer susceptibility, such as the breast cancer susceptibility loci *BRCA1* and *BRCA2*, neurofibromin 1 (*NF1*), retinoblastoma locus *RB1*, Wilms tumor locus *WT1*, and ataxia telangiectasia mutated (*ATM*). In total, > 8.6 Mb of baseline human reference sequence has been scanned across the 90 PDR samples, generating > 770 Mb of sequence (the equivalent of resequencing human chromosome 3 four times). Sixty percent of the candidate genes have been scanned for variation across > 75% of the entire reference gene sequence, and for many, nearly complete sequences are available. For each candidate, all exons, 1.5 kb upstream of the cDNA sequence, 1.5 kb downstream of the last exon, and a significant amount of intronic sequence have been examined. These efforts have uncovered > 50,000 SNPs and produced 4.5 million genotypes, which are cataloged at the GeneSNPs database (http://genome.utah.edu/genesnps) and the dbSNP database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp).

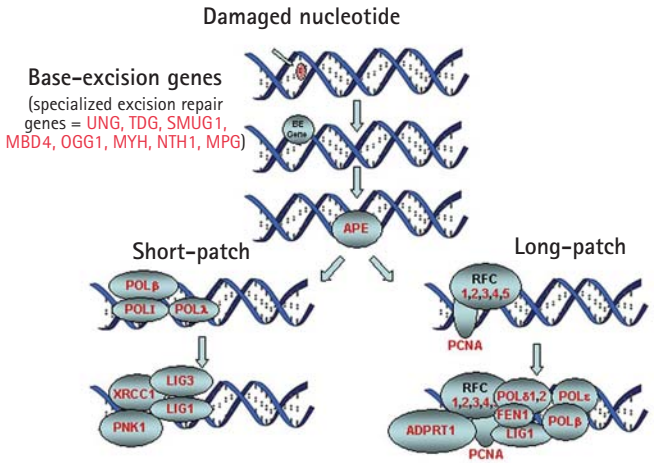

Figure 1. A schematic overview of the base-excision repair pathway adapted from Matsumoto (2001). This pathway is composed of more than 30 interacting genes. The polymorphism data across this pathways allow investigators to probe disease risk not only for individual members of the pathway but also for the entire pathway of interacting genes, thereby permitting analysis of multigenic contributions to disease risk.

Sequence Diversity in EGP Candidates

The overall nucleotide diversity (π) in the EGP genes is 6.7×10^{-4} (equivalent to one SNP every 1,500 bp between any two chromosomes), and the SNP frequency across the 180 chromosomes averaged one SNP every 173 bp. These figures are consistent with previous genomewide estimates of nucleotide diversity and SNP frequency (Carlson et al. 2004a; Halushka et al. 1999; Li and Sadler 1991; Nickerson et al. 1998; Sachidanandam et al. 2001; Stephens et al. 2001). However, significant variance around this mean is observed, and nucleotide diversity varied more than 63-fold, from 0.72×10^{-4} (equivalent to one SNP every 13,800 bp between any two chromosomes or a frequency of one SNP every 324 bp) for MARCKS-like protein (*MLP*) to 45.6×10^{-4} (one SNP every 221 bp between any two chromosomes, or a frequency of one SNP every 63 bp) for small proline-rich protein 1B (cornifin, *SPRRIB*). To contrast these genes, none of the 16 variable sites in *MLP* had an MAF > 5% in the sequenced population (n = 90 samples), and therefore, none can be considered common in the population. In comparison, 76 polymorphisms were identified in *SPRRIB*, and 43 of these variable sites (56%) were common in the population, having an MAF > 5%. The variability between these genes reveals the importance of detailed candidate gene studies. Although the average of all genes is consistent with genomewide levels of sequence diversity (Carlson et al. 2004b; Halushka et al. 1999; Li and Sadler 1991; Nickerson et al. 1998; Sachidanandam et al. 2001; Stephens et al. 2001), there is significant gene-to-gene and region-to-region variability that makes it difficult to predict the genetic structure of any given candidate or region in the genome (Clark et al. 2003). Because of this variability, there is only one gene sequenced to date, the cell cycle gene E2F transcription factor 2 (*E2F2*), that reflects the overall average diversity and size of the candidate genes sequenced by the EGP. A representation of the polymorphism distribution and gene structure of *E2F2* is shown in Figure 2, and each of the candidate genes being examined by the EGP is available in a similar format via

the GeneSNPs database (http://genome.utah.edu/genesnps). The GeneSNPs database was developed specifically for the EGP and is so highly regarded that many of its features have been emulated by other databases, including dbSNP (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp) and PharmGKB (http://www.pharmgkb.org/; Klein and Altman 2004).

For *E2F2*, 21.3 kb was scanned across the 90 PDR samples, and 112 single-nucleotide substitutions and six small insertion/deletion polymorphisms were identified. These polymorphisms are depicted by position in the gene in Figure 2 by vertical descending bars whose length is proportional to the allele frequency in the PDR. The nucleotide diversity across *E2F2* is 6.9×10^{-4}, or one SNP every 1.4 kb between two random chromosomes. The number of common polymorphisms is similar to that of other average genes, with 41% of the total (46 of 112 SNPs) having an MAF > 5% in the PDR 90 panel. Also typical of the average gene, *E2F2* has four coding SNPs (cSNPs), with two that are predicted to change the amino acid sequence (nonsynonymous) indicated by the red vertical bars in Figure 2.

Functional Analysis of the EGP SNPs

As described previously (Livingston et al. 2004), the average candidate gene contains approximately 34 common SNPs. Because functional analysis via genotype–phenotype studies or by animal models is costly, reducing the number of sites (from an average of 34) is a major consideration in effective study design. Two computational approaches have been taken by the EGP project to directly identify phenotypically relevant SNPs.

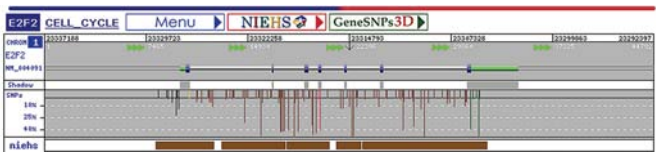

Figure 2. A GeneSNPs (http://www.genome.utah.edu/genesnps/) view of *E2F2*. *E2F2* represents an average of the genes scanned for polymorphism discovery based on its size and nucleotide diversity. The seven exons encoding this gene are depicted by blue rectangles for coding and green for untranslated (UTR) sequences in the mRNA. For this gene, 24 kb was scanned for polymorphisms, which includes sequences 5′ to the first exon (1.7 kb) and 3′ of the last exon (1 kb). Vertical descending lines indicate the position of the SNPs identified in this sequence. The length of the vertical lines represents the frequency of the minor allele, and the color indicates whether the SNP location is in flanking (black), intronic (brown), synonymous (yellow), nonsynonymous (red), or UTR (green) sequences.

One of these approaches has focused on testing the nonsynonymous (potentially functional) variations in coding sequences (Botstein and Risch 2003; Collins et al. 1997; Kruglyak and Nickerson 2001) for direct association studies and to target specific cSNPs for the development of new animal models. Of the nearly 50,000 SNPs found in the 371 candidate genes, 1,085 nonsynonymous cSNPs (ns-cSNPs) have been identified. Therefore, on average only 2% of the variability in a gene sequence is the result of amino acid substitutions.

Interestingly, despite an average of a little more than two ns-cSNPs per gene, there is substantial variability among the candidate genes, as shown in Figure 3. Of the 371 candidate genes sequenced to date, only 221 genes (60%) contained at least one ns-cSNP. Among genes with ns-cSNPs, there is substantial variation in the number of ns-cSNPs per gene, which ranges from 1 to 21. More than 15 ns-cSNPs were detected in five genes: insulin-like growth factor receptor 2 (*IGF2R*) with 21 ns-cSNPs, REV3-like, catalytic subunit of DNA polymerase zeta (*REV3L*) with 21 ns-cSNPs, protein kinase, DNA-activated, catalytic polypeptide (*PRKDC*) with 20 ns-cSNPs, exonuclease 1 (*EXO1*) with 17 ns-cSNPs, and the excision repair cross-complementing rodent repair deficiency, complementation group 6 (*ERCC6*) with 16 ns-cSNPs. Further analysis of these highly variable outlier genes could produce new functional insights and should be pursued in more detailed analyses.

In individual candidate genes, ns-cSNPs were further analyzed by applying two computational approaches that have been developed to detect functionally significant amino acid changes, SIFT (Ng and Henikoff 2003) and PolyPhen (Ramensky et al. 2002; Sunyaev et al. 2001). To date, 119 ns-cSNPs (~ 11% of total ns-cSNPs) have been identified as potentially deleterious by both of these approaches. Of these, only 11 potentially deleterious ns-cSNPs (dSNPs) had MAFs exceeding 5%. This important category of ns-cSNPs, the ones that commonly occur in the population, represents only a minor fraction of the total ns-cSNPs identified in the EGP (1% of the total ns-cSNPs and < 0.03% of the total SNPs identified). It is worth noting that the candidate genes with the highest number of ns-cSNPs also had multiple ns-cSNPs with predicted functional consequences based on both SIFT and PolyPhen predictions, including *REV3L*, *PRKDC*, and *EXO1*. *IGF2R* and *ERCC6* did not have sufficient comparative data for accurate prediction

with these two programs but are also likely to be highly polymorphic. A recent study exploring cSNPs in human genes associated with high-density cholesterol levels revealed larger numbers of rare ns-cSNPs with predicted functional significance when individuals at the extremes of the phenotypes were sequenced and compared (Cohen et al. 2004). Therefore, it is possible that perusing EGP genes with highly polymorphic coding regions will also be productive in terms of phenotype and direct functional analysis.

The vast majority of SNPs in the human genome are in noncoding sequences (> 91%). However, our ability to predict function in noncoding sequences is limited. Several new approaches are developing to predict functional regions in noncoding sequences through the application of comparative genomics and the mining of sequences that have been highly conserved through evolutionary history (Ahituv et al. 2004; Boffelli et al. 2004a, 2004b; Dieterich et al. 2003; Frazer et al. 2004; Sandelin et al. 2004). For the EGP, TraFaC (transcription factor binding site comparison; http://trafac.chmcc.org), a web-accessible tool for identifying transcription regulatory regions using a comparative sequence analysis approach, has been applied (Jegga et al. 2002). TraFaC generates a graphical output from BLASTZ alignments by comparing sequences of human and mouse orthologs or other sequences of interest. Potential transcription factor binding sites (TFBSs) are identified as conserved blocks in the two compared sequences. An example of the TraFaC output for cell division cycle 25a (*CDC25A*) is shown in Figure 4. This alignment illustrates SNPs in conserved consensus TFBSs, such as the variation (–503 C > T) in a putative *PAX1* (paired box gene 1) binding site in a gene with transcriptional activating properties. This potential *PAX1* binding site is located in the 5′-flanking sequence of *CDC25A* (Figure 4) and could adversely affect the regulation of gene expression.

Insights from comparative genomics are rapidly developing, and a number of approaches are being applied to search for functionally important non-cSNPs (Berjerano et al. 2004; Pennacchio and Rubin 2001, 2003; Thomas et al. 2003). It is likely that this area will continue to be a major focus in future studies of the gene function. In this regard, ongoing efforts from The ENCODE Project (ENCODE Consortium 2004), which is focused on rapidly determining the function of genomic sequences beyond the coding sequences, will likely be expanded to genes of interest for the EGP in the near future and could provide new candidate SNPs for functional analyses.

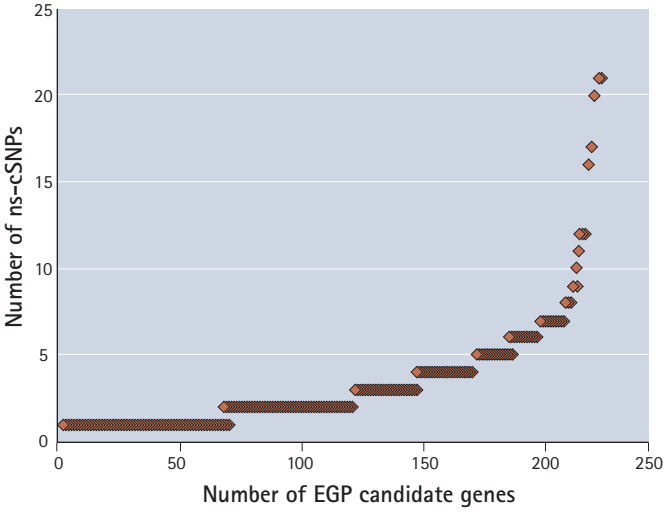

Figure 3. The number of ns-cSNPs per gene, for EGP candidate genes containing these substitutions. Of the 371 genes scanned across the 90 PDR samples, only 221 genes (60%) had one or more ns-cSNPs.

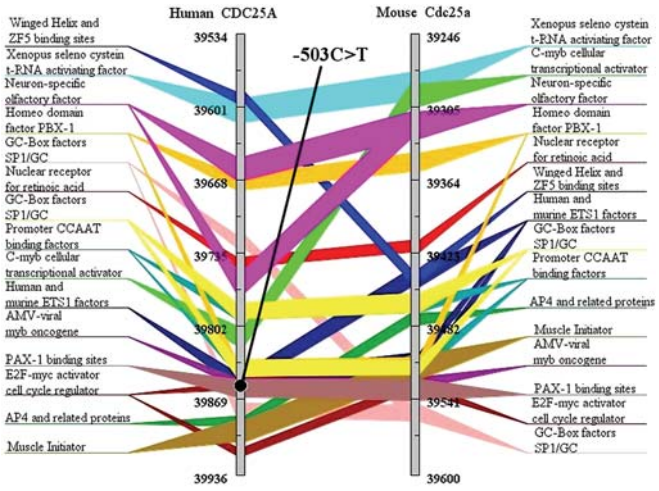

Figure 4. Conserved noncoding regions identified by TraFaC for *CDC25A*. The regulogram depicts shared *cis*-elements (vertical bars) between human (left) and mouse (right) sequences in the context of their sequence similarity. Identifying conserved mouse–human regions with consensus *cis*-regulatory elements and mapping non-cSNPs, TraFaC can be used to predict the potential adverse effects of polymorphisms on the regulation of gene and expression. The promoter region of human *CDC25A* and mouse *Cdc25a* reveals strong conservation of consensus TFBSs in relatively the same order of occurrence. The TFBSs occurring in both genes are highlighted as variously colored bars drawn across the two genes. An SNP identified in the promoter (–503 C > T) is highlighted.

Identifying New Functional Variation via Indirect Association Studies

Although new computational approaches to predict functional SNPs in the human genome are emerging, indirect association studies will ultimately be applied to identify SNPs with function that cannot be predicted *a priori* using approaches similar to those described above for SNPs in coding and noncoding sequences. Indirect association studies rely on linkage disequilibrium (LD) between genetic markers to measure the association between the SNPs genotyped, as well as the SNPs in LD with the assayed site and the disease phenotype (Collins et al. 1997). The number of sites required for genotyping any gene or region of the genome will greatly depend on the strength and extent of LD. For regions with strong LD and few haplotypes, only a few sites are required to represent or "tag" the gene or region. However, if the genomic region contains many haplotypes indicating low levels of LD, many more sites will be required for an association study of sufficient power. In the genes sequenced to date, there is much variability in the patterns of LD and common haplotype diversity (Figures 5 and 6). This is true across the human genome (Clark et al. 2003; Patil et al. 2001; Phillips et al. 2003; Reich et al. 2001; Stephens et al. 2001). Based on this, it is imperative to characterize the sequence and haplotype diversity of specific genomic regions of interest in human populations to rationally select SNPs for genotyping in an association study. In this respect, the EGP data set is providing new insights into these important questions in population genetics (Livingston et al. 2004; Wall and Pritchard 2003).

The gene-to-gene variability observed in the EGP for nucleotide diversity is also evident in site correlations or LD. Figure 5 illustrates some of the extremes observed in LD, as measured by the metric r^2, across the genes involved in environmental responses. For genes with average or high LD, such as BCL2/adenovirus E1B 19kDa interacting protein 1 (*BNIP1*) (Figure 5A) and *NF1* (Figure 5C), respectively, few sites are required for genotyping in association studies. However, for genes with very weak LD, such as cyclin D2 (*CCND2*; Figure 5B), many more sites will be required for a genetic association study because few sites within this gene are correlated. It is important to note that the extent of LD across a gene is independent of gene size. For example, LD extends across the 283-kb *NF1*, whereas fewer correlated sites are present in the smaller *CCND2*. For these genes with weak LD, attempts to choose sites with either LD-based (Carlson et al. 2004b) or haplotype-based (Johnson et al. 2001) selection will require typing a larger fraction of the common sites in the candidate gene. Although the stratified nature of the PDR can produce artifactual LD, the patterns of LD described here represent the range of observed patterns within the EGP data set. Particularly for genes that exhibit strong LD (e.g., *NF1*), these patterns appear to be consistent among the ethnic subpopulations in the PDR.

Although associations between individual sites and phenotype have proven useful in uncovering associations in the human genome (Meirhaeghe and Amouyel 2004; Tempfer et al. 2004), it is clear that the interactions between multiple sites within a region or gene may also be important and can be detected via haplotype associations. The best example of this is the association between a haplotype in the apolipoprotein E gene (*APOE4*) and Alzheimer's disease (Corder et al. 1993). There are many factors that influence haplotype structure, including the mutation, gene conversion, and recombination rates. Figure 6 shows the distribution in the haplotypes per gene for 330 genes from the EGP. Haplotypes were only inferred from common SNPs with > 5% MAF and ranged from 2 for *FAU* [Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed (fox derived); ribosomal protein S30] to 175 for *IGF1R* (insulin growth factor 1 receptor). It is worth noting that the average number of haplotypes per gene is 38 for this data set. However, haplotype diversity will be greatly influenced by recombination, and recent reports suggest that genes with extreme haplotype diversity may contain one or more hotspots of recombination that would greatly increase the overall number of haplotypes for any given candidate (Crawford et al. 2004a, 2004b), and several approaches have emerged to identify hotspots of recombination and will aid in developing new approaches to haplotype tagging for association analysis (Crawford et al. 2004a; McVean et al. 2004).

Future Prospects for the EGP

As envisioned by Dr. Kenneth Olden in 1997, the EGP has generated significant new insights into the diversity and genetics of environmental response genes. The initial set of target genes, 550 altogether, will be

completed over the next 6 months, and targets for the next phase are already under development. It is likely that DNA sequencing will play an increasingly important role in the EGP and in all future genetic analysis with its decreasing costs and rapidly expanding scale. Over the next decade, new approaches in *in situ* sequencing will be tested. If successful, it is likely that such genome-based resequencing will emerge as a dominant genotyping strategy as well (Collins et al. 2003; Shendure et al. 2004). Indeed, if the $1,000 per resequenced human genome becomes a reality, sequence analysis will play an increasingly important role in whole-genome association studies in human populations because the entire spectrum of variation (common and rare) can be uncovered in a single pass.

Until complete resequencing becomes a reality, only more limited subsets of the variation identified across the human genome will be applied in association studies. Several approaches are being taken to reduce the number of sites to be tested in human studies. The first

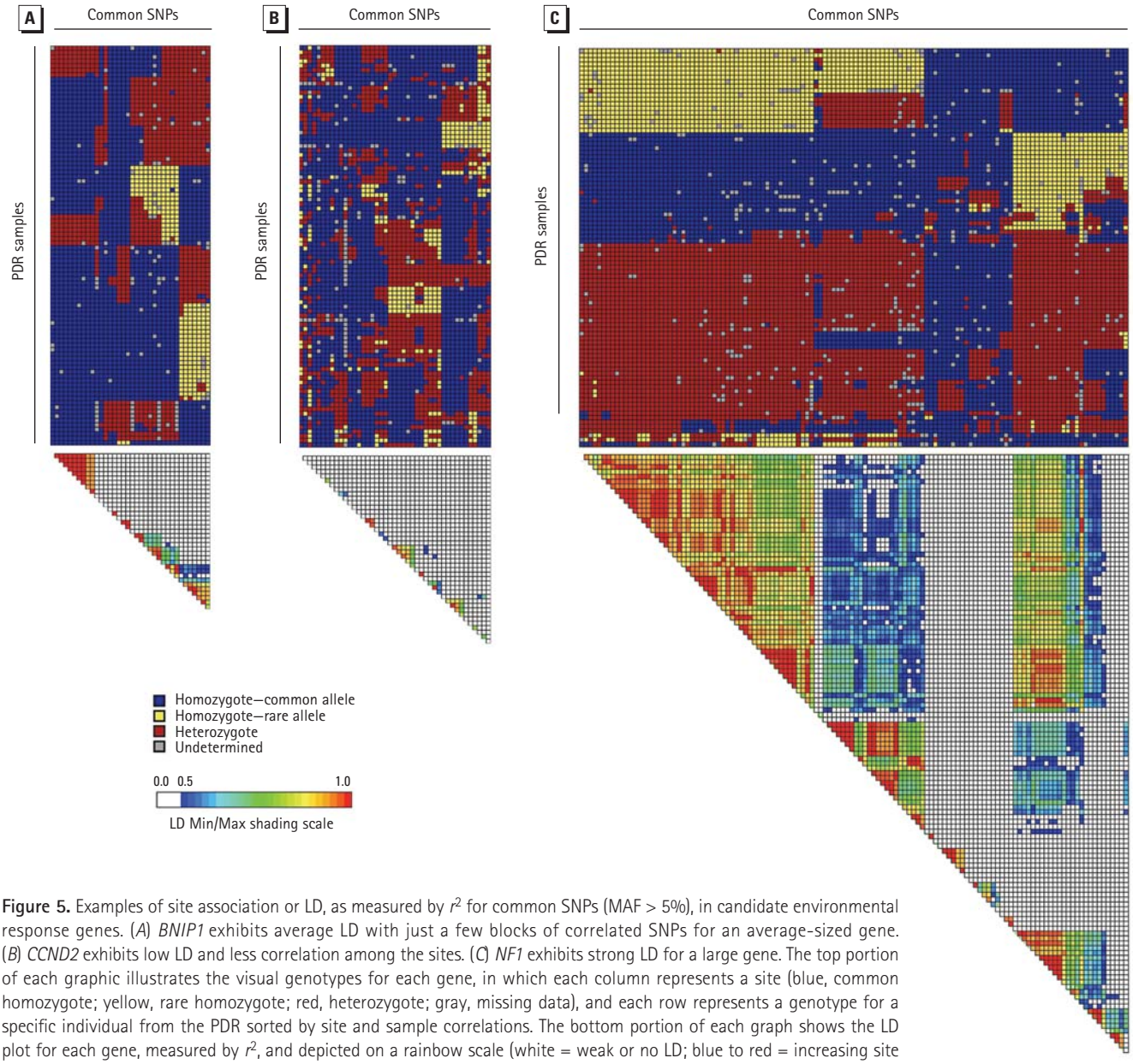

Figure 5. Examples of site association or LD, as measured by r^2 for common SNPs (MAF > 5%), in candidate environmental response genes. (*A*) *BNIP1* exhibits average LD with just a few blocks of correlated SNPs for an average-sized gene. (*B*) *CCND2* exhibits low LD and less correlation among the sites. (*C*) *NF1* exhibits strong LD for a large gene. The top portion of each graphic illustrates the visual genotypes for each gene, in which each column represents a site (blue, common homozygote; yellow, rare homozygote; red, heterozygote; gray, missing data), and each row represents a genotype for a specific individual from the PDR sorted by site and sample correlations. The bottom portion of each graph shows the LD plot for each gene, measured by r^2, and depicted on a rainbow scale (white = weak or no LD; blue to red = increasing site association strength or LD).

approach is to explore the sites associated with coding and conserved noncoding sequences that can be functionally assessed via computational methods.

cSNPs represents only a small subset of the variation across the genome, but if common diseases arise from mechanisms similar to rare Mendelian diseases, this subset of SNPs should be tested. However, arguments for and against this hypothesis have been raised, and numerous reviews of the issues involved are available (Cardon and Bell 2001; Pritchard and Cox 2002; Reich and Lander 2001). By extrapolating current findings for the EGP across the human genome for an estimated 24,000 genes (Ewing and Green 2000; International Human Genome Sequencing Consortium 2004; Lander et al. 2001), 150,000 cSNPs (with an MAF > 1%) may be identified. This figure is similar to prior predictions (Kruglyak and Nickerson 2001). On average, 50% of the cSNPs identified will lead to amino acid substitutions. Therefore, 75,000 amino acid–altering SNPs may be identified across the human genome. Using new computational approaches such as SIFT and PolyPhen to score potential functionally ns-cSNPs, the subset of cSNPs could drop to approximately 8,250 cSNPs with predicted functional importance. However, only a small fraction of these cSNPs, approximately 1,000, will have MAFs > 5% and could be easily tested with newly developed low-cost genotyping strategies on adequately sized human population cohorts. Although this estimate could potentially reflect the relatively high conservation bias of the EGP genes, the gene-to-gene variation observed for the EGP is consistent with previous observations of sets of genes involved in inflammation, lipid metabolism, and endocrine function (Cargill et al. 1999; Carlson et al. 2004a; Crawford et al. 2004b; Halushka et al. 1999; Stephens et al. 2001).

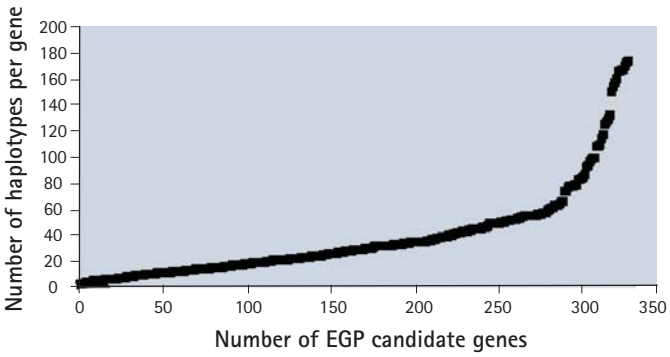

Figure 6. Haplotypes diversity across EGP candidate genes. Haplotypes were inferred using PHASE software, version 2.1 (Stephens and Donnelly 2003) for all SNPs with an MAF ≥ 5% in the EGP samples of 90 individuals representing 180 chromosomes.

It is also possible to predict SNPs in noncoding sequences with functional significance using comparative genomic approaches (Dieterich et al. 2003; Frazer et al. 2004; Jegga et al. 2002; Sandelin et al. 2004; Schwartz et al. 2003). Furthermore, data from the ENCODE Project (ENCODE Consortium 2004) should generate new paradigms for these analyses. The development of new comparative sequence data from divergent species in the evolutionary tree (Wallis et al. 2004) will greatly enhance studies to identify SNP subsets in noncoding regions with predicted function (Bejerano et al. 2004). Even when more refined approaches develop for noncoding regions of biological interest, the sizes of the emerging SNP panels are likely to be comparable with cSNP sets.

Several of the predicted functional SNPs from the EGP are being developed by the Comparative Mouse Genomics Centers Consortium (CMGCC; NIEHS 2005) to generate transgenic and knockout mouse models based on human DNA sequence variants identified in environmentally responsive genes. These mouse models will become tools to improve our understanding of the biological significance of human DNA polymorphism, and many of the computationally mined SNPs from coding sequences are being translated into appropriate animal models for further studies (Ladiges et al. 2004; NIEHS 2005). Initially, the CMGCC is focusing on variation in genes involved in DNA repair or cell cycle control, because many of these are well-characterized, environmentally responsive genes that can be translated into many current studies (Angus et al. 2003; Bahassi el et al. 2002).

According to the common disease/common variant hypothesis, the genetic risk factors underlying common diseases are likely common, modest risk alleles in the human population (Altshuler et al. 2000a; Halushka et al. 1999; Reich and Lander 2001). These alleles may not be adequately predicted by current computational tools, which are better suited for identifying the highly penetrant alleles associated with rare Mendelian diseases. These denser SNP sets, like those now available for the EGP project, will engender a second, indirect approach to association mapping. This approach will develop SNP sets to exploit LD (or SNP associations) to capture the sites that cannot be

predicted by computational tools to be functionally important.

It is worth noting that our limited ability to predict SNPs with functional consequences has led to the development of several other large-scale projects to discover and type common polymorphisms across the human genome (International HapMap Consortium 2003; Patil et al. 2001). These are far less comprehensive then the current EGP project but more global in their approach to the genome. An initial set of 600,000 SNPs is being developed, but even this set is under expansion because of the wide variation in LD or SNP association known to exist across the human genome. In fact, the complete data sets available through the EGP have been one driver of the need to increase the density of the current HapMap. However, the global views provided by these genomewide data sets are important, and there is an effort under way to type the common variation identified through the EGP on the HapMap samples so that the data from the EGP genes are integrated with resources emerging for the entire human genome.

From its inception, the EGP has provided the genotype data needed to drive the next-generation association mapping of genotype–phenotype–environmental interactions. This project has revealed the broad gene-to-gene variation in sequence diversity, LD, and haplotype diversity present in the human genome. Additionally, the EGP is one of the first projects to explore large noncoding genic sequences in the human genome and, as such, is one of the few projects that can fully inform association studies of candidate genes or association analysis of entire gene pathways. With the application of even a small pathway of genes, such as the base-excision repair pathway, it is clear that a new generation of software tools for association mapping will be required to fully explore even a simple candidate gene pathway for SNPs, haplotypes, and interactions between genes and with environmental modifiers. This is a key frontier for the EGP and is likely to be addressed during the next phase of the project.

Summary

To explore the relationship between environmental exposure and genetic susceptibility in the etiology of common diseases, the Environmental Genome Project is scanning environmental response genes involved in DNA repair, cell cycle control, apoptosis, drug metabolism, and other pathways for single nucleotide polymorphisms (SNPs). To date, more than 370 candidate environmental response genes have been examined and 50,000 SNPs identified across 8.6 Mb of baseline sequence, providing valuable resources for association mapping of genotype–phenotype–environment interactions. Additionally, these data are stimulating ongoing efforts to develop mouse models of potentially functional polymorphisms; explore the common variant/common disease hypothesis; address the ethical, legal, and social implications of the genetics of environmental disease susceptibility; and develop new tools, views, and strategies to improve the discovery of genetic variations responsible for sensitivity to environmental agents.
doi:10.1289/ehp.7922 available via *http://dx.doi.org/*

Notes

Address correspondence to D.A. Nickerson, Department of Genome Sciences, University of Washington, Box 357730, Seattle, WA 98195-7730 USA. Telephone: (206) 685-7387. Fax: (206) 221-6498. E-mail: debnick@u.washington.edu

The authors declare they have no competing financial interests.

References

Ahituv N, Rubin EM, Nobrega MA. 2004. Exploiting human–fish genome comparisons for deciphering gene regulation. Hum Mol Genet 13(spec no 2):R261–R266.

Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, et al. 2000a. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet 26:76–80.

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, et al. 2000b. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407:513–516.

Angus SP, Solomon DA, Kuschel L, Hennigan RF, Knudsen ES. 2003. Retinoblastoma tumor suppressor: analyses of dynamic behavior in living cells reveal multiple modes of regulation. Mol Cell Biol 23:8172–8188.

Bahassi el M, Conn CW, Myer DL, Hennigan RF, McGowan CH, Sanchez Y, et al. 2002. Mammalian Polo-like kinase 3 (Plk3) is a multifunctional protein involved in stress response pathways. Oncogene 21: 6633–6640.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, et al. 2004. Ultraconserved elements in the human genome. Science 304:1321–1325.

Belinsky SA. 2004. Gene-promoter hypermethylation as a biomarker in lung cancer. Nat Rev Cancer 4:707–717.

Bell DA, Taylor JA. 1997. Genetic analysis of complex disease. Science 275:1327–1328; author reply 1329–1330.

Berwick M. 2000. Gene-environment interaction in melanoma. Forum (Genova) 10:191–200.

Boffelli D, Nobrega MA, Rubin EM. 2004a. Comparative genomics at the vertebrate extremes. Nat Rev Genet 5:456–465.

Boffelli D, Weer CV, Weng L, Lewis KD, Shoukry MI, Pachter L, et al. 2004b. Intraspecies sequence comparisons for annotating genomes. Genome Res 14:2406–2411.

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.

Cardon LR, Bell JI. 2001. Association study designs for complex diseases. Nat Rev Genet 2:91–99.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, et al. 1999.

Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231–238.

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. 2004a. Mapping complex disease loci in whole-genome association studies. Nature 429:446–452.

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. 2004b. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106–120.

Clark AG, Nielsen R, Signorovitch J, Matise TC, Glanowski S, Heil J, 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.

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.

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

Collins FS, Green ED, Guttmacher AE, Guyer MS. 2003. A vision for the future of genomics research. Nature 422:835–847.

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

Cooper DN, Smith BA, Cooke HJ, Niemann S, Schmidtke J. 1985. An estimate of unique DNA sequence heterozygosity in the human genome. Hum Genet 69:201–205.

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, et al. 1993. Gene dose of apoliprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261:921–923.

Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, et al. 2004a. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet 36:700–706.

Crawford DC, Carlson CS, Rieder MJ, Carrington DP, Yi Q, Smith JD, et al. 2004b. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. Am J Hum Genet 74:610–622.

Dahlqvist A, Hammond JB, Crane RK, Dunphy JV, Littman A. 1963. Intestinal lactase deficiency and lactose intolerance in adults. Preliminary report. Gastroenterology 45:488–491.

Dieterich C, Wang H, Rateitschak K, Luz H, Vingron M. 2003. CORG: a database for COmparative Regulatory Genomics. Nucleic Acids Res 31:55–57.

Doll R. 1975. Pott and the path to prevention. Arch Geschwulstforsch 45:521–531.

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306:636–640.

Eriksson S. 1965. Studies in alpha 1-antitrypsin deficiency. Acta Med Scand (suppl) 432:1–85.

Ewing B, Green P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. Nat Genet 25:232–234.

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. 2004. VISTA: computational tools for comparative genomics. Nucleic Acids Res 32:W273–W279.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Haemmerli UP, Kistler H, Ammann R, Marthaler T, Semenza G, Auricchio S, et al. 1965. Acquired milk intolerance in the adult caused by lactose malabsorption due to a selective deficiency of intestinal lactase activity. Am J Med 38:7–30.

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

Han J, Colditz GA, Samson LD, Hunter DJ. 2004. Polymorphisms in DNA double-strand break repair genes and skin cancer risk. Cancer Res 64:3009–3013.

Hayward NK. 2003. Genetics of melanoma predisposition. Oncogene 22:3053–3062.

Hegele RA. 1997. Candidate genes, small effects, and the prediction of atherosclerosis. Crit Rev Clin Lab Sci 34:343–367.

Ide H, Kotera M. 2004. Human DNA glycosylases involved in the repair of oxidatively damaged DNA. Biol Pharm Bull 27:480–485.

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

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431:931–945.

Jegga AG, Sherwood SP, Carman JW, Pinski AT, Phillips JL, Pestian JP, et al. 2002. Detection and visualization of compositionally similar cis-regulatory element clusters in orthologous and coordinately controlled genes. Genome Res 12:1408–1417.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, et al. 2001. Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237.

Kelada SN, Stapleton PL, Farin FM, Bammler TK, Eaton DL, Smith-Weller T, et al. 2003. Glutathione *S*-transferase M1, T1, and P1 polymorphisms and Parkinson's disease. Neurosci Lett 337:5–8.

Klein TE, Altman RB. 2004. PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base [Editorial]. Pharmacogenomics J 4:1.

Klotz AP. 1964. Intestinal lactase deficiency and diarrhea in adults. Am J Dig Dis 10:345–354.

Kruglyak L. 1997. The use of a genetic map of biallelic markers in linkage studies. Nat Genet 17:21–24.

Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–144.

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

Ladiges W, Kemp C, Packenham J, Velazquez J. 2004. Human gene variation: from SNPs to phenotypes. Mutat Res 545:131–139.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lander ES, Schork NJ. 1994. Genetic dissection of complex traits [published erratum Science 266:353] . Science 265:2037–2048.

Li WH, Sadler LA. 1991. Low nucleotide diversity in man. Genetics 129:513–523.

Lieberman J, Mittman C, Schneider AS. 1969. Screening for homozygous and heterozygous alpha 1-antitrypsin deficiency. Protein electrophoresis on cellulose acetate membranes. JAMA 210:2055–2060.

Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, et al. 2004. Pattern of sequence variation across 213 environmental response genes. Genome Res 14:1821–1831.

Matsumoto Y. 2001. Molecular mechanism of PCNA-dependent base excision repair. Prog Nucleic Acid Res Mol Biol 68:129–138.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584.

Meirhaeghe A, Amouyel P. 2004. Impact of genetic variation of PPARgamma in humans. Mol Genet Metab 83:93–102.

Moffatt MF, Cookson WO. 1999. Genetics of asthma and inflammation: the status. Curr Opin Immunol 11:606–609.

Mohrenweiser HW, Xi T, Vazquez-Matias J, Jones IM. 2002. Identification of 127 amino acid substitution variants in screening 37 DNA repair genes in humans. Cancer Epidemiol Biomarkers Prev 11:1054–1064.

Motulsky AG. 1972. Hemolysis in glucose-6-phosphate dehydrogenase deficiency. Fed Proc 31:1286–1292.

Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814.
Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, et al. 1998. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat Genet 19:233–240.
NIEHS. Comparative Mouse Genomics Centers Consortium. Research Triangle Park, NC:National Institute of Environmental Health Sciences. Available: http://www.niehs.nih.gov/cmgcc/pub.htm [accessed 8 February 2005].
Olden K, Wilson S. 2000. Environmental health and genomics: visions and implications. Nat Rev Genet 1:149–153.
Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, et al. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723.
Pennacchio LA, Rubin EM. 2001. Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2:100–109.
Pennacchio LA, Rubin EM. 2003. Comparative genomic tools and databases: providing insights into the human genome. J Clin Invest 111:1099–1106.
Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, et al. 2003. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet 33:382–387.
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 30:3894–900.
Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, et al. 2001. Linkage disequilibrium in the human genome. Nature 411:199–204.
Reich DE, Gabriel SB, Altshuler D. 2003. Quality and completeness of SNP databases. Nat Genet 33:457–458.
Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. Trends Genet 17:502–510.
Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273:1516–157.
Rothman N, Garcia-Closas M, Stewart WT, Lubin J. 1999. The impact of misclassification in case-control studies of gene-environment interactions. IARC Sci Publ 148:89–96.
Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933.
Sandelin A, Wasserman WW, Lenhard B. 2004. ConSite: web-based prediction of regulatory elements using cross-species comparison. Nucleic Acids Res 32:W249–W252.
Schwartz S, Elnitski L, Li M, Weirauch M, Riemer C, Smit A, et al. 2003. MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. Nucleic Acids Res 31:3518–3524.
Shendure J, Mitra RD, Varma C, Church GM. 2004. Advanced sequencing technologies: methods and goals. Nat Rev Genet 5:335–344.
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 JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, et al. 2001. Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489–493.
Sunyaev S, Ramensky V, Koch I, Lathe W III, Kondrashov AS, Bork P. 2001. Prediction of deleterious human alleles. Hum Mol Genet 10:591–597.
Tempfer CB, Schneeberger C, Huber JC. 2004. Applications of polymorphisms and pharmacogenomics in obstetrics and gynecology. Pharmacogenomics 5:57–65.
Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424:788–793.
Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. 2001. The sequence of the human genome. Science 291:1304–1351.
Wall JD, Pritchard JK. 2003. Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 4:587–597.
Wallis JW, Aerts J, Groenen MA, Crooijmans RP, Layman D, Graves TA, et al. 2004. A physical map of the chicken genome. Nature 432:761–764.
Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, et al. 1998. Large-scale identification, mapping, and genotyping of single nucleotide polymorphisms in the human genome. Science 280:1077–1082.
Wang WY, Todd JA. 2003. The usefulness of different density SNP maps for disease association studies of common variants. Hum Mol Genet 2:3145-3149.
Waters MD, Fostel JM. 2004. Toxicogenomics and systems toxicology: aims and prospects. Nat Rev Genet 5:936–948.