Final Report - DE-FG03-97ER62385

Principal Investigator: Deborah A. Nickerson, Department of Molecular Biotechnology, University of Washington

Single nucleotide polymorphism (SNPs) are the most common form of sequence variation in the human genome, and based on their natural frequency, they are also likely to be the underlying cause of most phenotypic differences in humans. Since SNPs are found in both coding and non-coding regions of the genome, randomly distributed markers as well as markers clustered in genes can be discovered. The majority of SNPs found in coding regions (cSNPs) are single base substitutions that may or may not lead to amino acid substitutions. Some cSNPs can result in changes in the activity of protein by altering a functionally important amino acid residue(s), and these are of interest for their potential links with phenotype, e.g. disease susceptibility or resistance. Other cSNPs (synonymous and non-synonymous) may prove useful for their potential links to functional cSNPs via linkage disequilibrium mapping.

With funding from the Department of Energy, we have explored approaches: (1) to improve technology to find SNPs in the human genome using fluorescence-based sequencing, and (2) to mine SNPs from genome resources such as expressed-sequence tag (EST) sequences.

(1) Improving the identification of DNA variations by fluorescence-based sequencing: Direct sequencing of PCR products is one of the most automated approaches for scanning the genome for DNA polymorphisms and mutations. We have evaluated methods to improve the accuracy of identifying DNA polymorphisms/mutations using fluorescence-based sequencing. We have developed several programs that aid in the detection of single nucleotide polymorphisms - one that detects polymorphisms even in just as heterozygotes (two bases at one location) among homozygous sequences. Funding from the DOE aided particularly in improving the accuracy of genotyping efficiency of the PolyPhred program which works together with Phred, Phrap and Consed programs already in use for large-scale sequencing projects. PolyPhred is use by more than 300 laboratories world-wide and applied routinely for polymorphism detection. We also developed a tool known as RefComp which aids in the detection of high quality mismatches or mutations in a sequence compared to reference sequence. This program has proven quite effective in the detection of mutation in human mitochondrial sequences.

Publications on SNP detection:

Nickerson, D.A., Tobe, V.O., and Taylor, S.L. 1997. PolyPhred: Detecting and genotyping single nucleotide substitutions by fluorescence-based resequencing, Nucleic Acids Research, 25: 2745-2751.

Rieder, M.J., Taylor, S.L., Tobe, V.O. and Nickerson, D.A. 1998. Automating the Identification of DNA Variations using Quality-Based Fluorescence Resequencing: Analysis of the Human Mitochondrial Genome, Nucleic Acid Research 26: 967-973.

(2) To mine SNPs from genome resources such as expressed-sequence tag (EST) sequences: We have investigated the assembly ESTs to develop a new approach to finding SNPs in the coding regions of human genes and new potential mutations in human DNA. For this analysis, ESTs representing the full-length sequence of 850 human genes were identified and 201 candidate cSNPs found. Of these 87 cSNPs were predicted to lead to amino acid changes. To identify cSNPs, we developed a heuristic algorithm that uses Phred quality scores to detect candidate SNPs. Phred sequence quality (q) is related to the error probability of each base-call via the transformation $q = -10\log(err)$, where err is the error probability. In these sequences, candidate cSNPs with Phred quality > 20 (< 1% error probability) at the site of the mismatch and whose immediately surrounding sequence had Phred quality > 20 were identified. After filtering, 223 candidate cSNPs were detected. Visual inspection of the traces revealed two types of systematic errors, resulting from base calling (n=21) and misalignment (n=1) problems. To improve this strategy, we developed an approach with a modification of PolyPhred known as RefComp to detect homozygous mismatches in overlapping cloned sources of cDNA or genomic DNA. All of the SNPs identified by our studies have submitted to dbSNP.

Publications on EST SNPs:

Garg, K., Green, P. and Nickerson, D.A. 1999. Identifying candidate coding region single nucleotide polymorphisms (cSNPs) using assembled expressed sequence tags (ESTs), Genome Res. 9: 1087-1092.